Supplementary Material for Learned Spatial Representations for Few-shot Talking-Head Synthesis

Moustafa Meshry Saksham Suri Larry S. Davis Abhinav Shrivastava

University of Maryland, College Park

A. Appendix

A.1. Implementation details

Dataset pre-processing. The released VoxCeleb2 dataset [1] contains pre-processed videos to have a center crop around the face. We uniformly sample 10 frames from each video and obtain the facial landmarks using an off-the-shelf facial landmarks detector [2]. Once the landmarks are obtained we use the same procedure as [3] to connect the facial landmarks to obtain contours for different face parts (*e.g.,* eyes, nose, lips . . . etc.). We observe that the facial landmarks extraction fails for a small fraction of videos, which we opted to ignore. We also segment each frame using the face parsing tool provided by [4] to obtain the oracle segmentation maps for pre-training the layout prediction network. The face parsing network performs poorly on VoxCeleb2 frames due to the domain gap, in terms of image resolution and the distribution of head poses, between the datasets used to train the face parsing network [4], and the cropped VoxCeleb2 videos. We observe that the face segmentation network [4] better captures different details at different resolutions. For example the segmentation result at the original VoxCeleb2 resolution better captures larger regions like the hair, neck and clothes. On the other hand, upsampling the frame to the resolution used for training the segmentation network [4] gives better segmentation results for the finer and smaller regions like the nose, eyes, mouth, and ears. So, to improve the oracle segmentations, we segment each frame twice at 256x256 and 512x512 resolutions and merge the coarse and fine semantic classes from both results.

Encoder networks. We use a resnet encoder for both the layout and style encoders $\{E^l, E^s\}$. The encoder architecture has 5 downsampling blocks, followed by a fully connected layer that generates a 512-dimensional latent code. The architecture for the residual blocks is borrowed from [5], with replacing *average-pooling* with *blur-pooling*. We use 32 feature maps at the first encoder layer and double this number after each downsampling block with a maximum of 512 feature maps. We follow [3] and concatenate the facial landmarks to the few-shot RGB images before feeding them to the encoder.

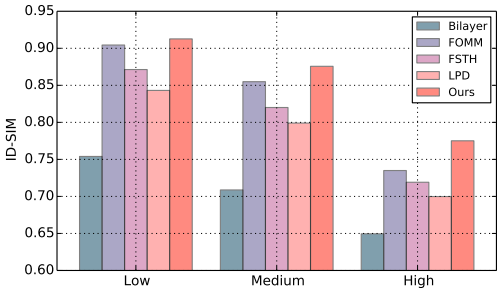

Figure 1: Identity similarity metric (ID-SIM) for the single-shot setting across three test subsets representing low, medium and high variance between the source and target poses. The performance gap widens in our favor as the pose variance increases.

Layout generator. We use a traditional UNet architecture [6] with residual blocks. The residual blocks are borrowed from [5] with replacing *BatchNorm* with *Instance Norm* and applying adaptive instance normalization (*AdaIN*) [7]. The smallest and largest number of feature maps are 32 and 512 respectively, and we use *blur-pool* and *bilinear* upsampling in the downsampling and upsampling blocks respectively.

Image generator. We use a SPADE generator architecture [8] with replacing *BatchNorm* with *Instance Norm*. We also use 32 feature maps at the last generator layer and 64 feature maps in each SPADE block, compared to 64 and 128 feature maps respectively in the original architecture [8]. The input to each SPADE block is the concatenation of the predicted layout map and the facial landmarks.

Discriminator. We borrow the architecture of the discriminator network from [9], with reducing the smallest number of feature maps from 64 to 32. We also use a non-saturating logistic loss with gradient penalty [10].

Training. We follow [11] and use equalized learning rate in all of our networks. We pre-train the layout prediction

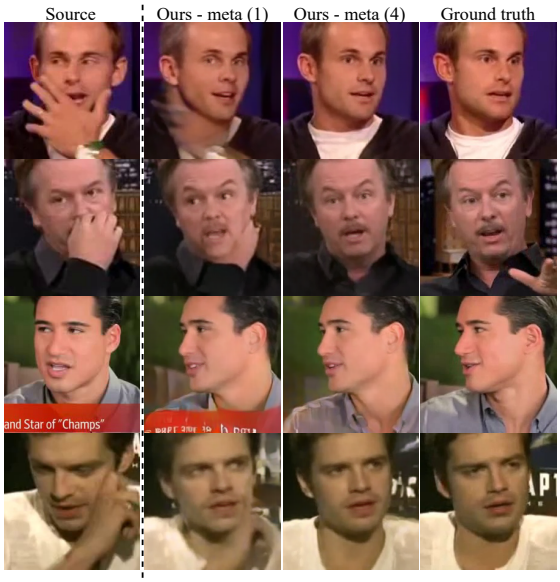

Figure 2: Averaging latent codes from multi-shot inputs successfully filters out transient occluders and maintains only the desired information for novel view head synthesis.

network for 2 epochs, followed by training the full pipeline for 8 epochs. Our best model was left to train for an extra 5 epochs, which mainly improves the FID score, while slightly improving the other quantitative metrics as well. We use an *Adam* optimizer [12] with $\beta_1 = 0, \beta_2 = 0.999$, and a learning rate of 0.001 for all networks. We linearly decay the learning rate by a factor of 100 during the last epoch. For more implementation and training details, we will release the code and training scripts.

A.2. Robustness to pose variation

Here we investigate the robustness of different methods against the pose variation between the source and the target images. First, we cluster the test set into low, medium and high pose variance based on the mean normalized keypoint difference (NMKE) between the source and target ground truth images. Then we compute the identity similarity metric (ID-SIM) per each cluster for the single-shot setting and report the results in figure 1. The performance gap between our method and the baselines widens as the pose variance increases, indicating that our method has better robustness against pose variation. Note that we report the results only for the single-shot setting, where the performance gap with the FOMM baseline [13] is close. However, our method significantly outperforms FOMM in the multi-shot setting, as we show in Section A.4.1.

A.3. Effect of latent averaging

Given K-shot inputs, we follow [3] and obtain a single layout and style latents $\{z^l, z^s\}$ by averaging the K layout and style latents computed from the inputs respectively. We observe that averaging the K latents cancels out view-specific information and transient occluders, and successfully maintains the implicit 3D information needed for novel view head synthesis. Figure 2 shows some examples that highlight this effect. The single-shot source images show some transient occluders like the subjects' hand or news bar, which in turn corrupts our single-shot output. However, increasing the inputs to four frames successfully filters out the transient occluders and results in clean outputs.

Table 1: Comparison with the FOMM baseline [13]. While FOMM cannot benefit from multiple input frames, our method shows a significant improvement over FOMM with as few as 4-shot inputs.

Method	PSNR↑	SSIM↑	LPIPS↓	ID-SIM↑	NMKE↓	FID↓
FOMM [13]	**18.20**	**0.635**	0.236	0.869	0.061	56.10
Ours-meta (K=1)	17.27	0.598	0.241	0.869	0.041	48.11
Ours-ft (K=1)	17.37	0.605	**0.232**	**0.886**	**0.041**	**45.69**
Ours-meta (K=4)	18.90	0.638	0.192	0.909	0.039	43.19
Ours-ft (K=4)	**19.33**	**0.661**	**0.171**	**0.930**	**0.037**	**34.31**

A.4. More comparative results

A.4.1 Comparison with FOMM

The FOMM baseline [13] accurately reconstructs the background and other static regions due to its warping-based nature. Therefore, it achieves lower reconstruction error (PSNR and SSIM) than our approach in the single-shot setting, even if their output contains clear artifacts in the face area. However, one limitation to FOMM is that it cannot utilize more input frames to its advantage. On the other hand, Table 1 shows that our approach benefits from as few as four input frames to outperform FOMM, even in the meta-learned mode. Subject fine-tuning further improves our performance to outperform FOMM by a wide margin in all metrics.

A.4.2 More comparative evaluation

We report the quantitative details for the effect of increasing the number of K-shot inputs, as well as the effect of subject fine-tuning in Table 2. We observe similar conclusions to those obtained from Figure 6 in the main paper. LPD [14] performs very poorly in the meta-learned setting, and only outperforms the FSTH baseline [3] in the subject fine-tuning setting. On the other hand, our method consistently outperforms the baselines in all metrics across different settings. Furthermore, the performance of our method at $K = 4$ is on-par with or outperforms the baselines evaluated at $K = 32$ across all metrics. Since the LPD [14] baseline does not predict the background and re-crops the input/output frames, we subtract the background and compare with their corresponding cropped ground truths for quantitative analysis. We also exclude LPD from frame reconstruction evaluation since its outputs do not align with the rest of the methods.

Table 2: Detailed quantitative comparison with the few-shot baselines, showing the effect of both increasing the K-shot inputs and subject-specific fine-tuning.

K	Method	No Subject Fine-tuning						Subject Fine-tuned					
		PSNR↑	SSIM↑	LPIPS↓	ID-SIM↑	NMKE↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	ID-SIM↑	NMKE↓	FID↓
1	FSTH	16.80	0.570	0.259	0.801	0.048	51.12	16.92	0.597	0.263	0.836	0.049	53.07
	LPD	–	–	–	0.732	0.072	80.20	–	–	–	0.837	0.070	48.48
	Ours	**17.27**	**0.598**	**0.241**	**0.869**	**0.041**	**48.11**	**17.37**	**0.605**	**0.232**	**0.886**	**0.041**	**45.69**
4	FSTH	–	–	–	–	–	–	–	–	–	–	–	–
	LPD	–	–	–	0.755	0.069	79.67	–	–	–	0.909	0.058	38.81
	Ours	**18.90**	**0.638**	**0.192**	**0.909**	**0.039**	**43.19**	**19.33**	**0.661**	**0.171**	**0.930**	**0.037**	**34.31**
8	FSTH	17.86	0.600	0.225	0.836	0.046	46.38	18.35	0.647	0.218	0.899	0.044	45.15
	LPD	–	–	–	0.760	0.068	77.00	–	–	–	0.922	0.056	35.87
	Ours	**19.18**	**0.645**	**0.186**	**0.917**	**0.039**	**44.23**	**19.65**	**0.675**	**0.160**	**0.940**	**0.036**	**32.54**
32	FSTH	18.66	0.613	0.207	0.843	0.044	44.85	19.69	0.686	0.171	0.927	0.041	33.69
	LPD	–	–	–	0.769	0.066	63.47	–	–	–	0.935	0.054	33.96
	Ours	**19.35**	**0.650**	**0.182**	**0.921**	**0.037**	**43.32**	**19.98**	**0.690**	**0.146**	**0.948**	**0.038**	**28.26**

Figure 3: Extending Figure 5 of the main paper. More results comparing our performance to the few-shot baselines with respect to increasing the the K-shot inputs and applying subject fine-tuning.

Also, the authors of FSTH [3] only provide their output for $K = \{1, 8, 32\}$ and they did not release their code. Therefore, we don't report their performance for $K = 4$.

A.4.3 More qualitative comparisons

Here, we expand our qualitative comparisons of the main paper in both the multi-shot and single-shot settings. Figure 3 extends Figure 5 of the main paper. It shows more examples comparing the effect of increasing the K-shot inputs and applying subject fine-tuning between our method and the baselines. Figure 4 shows more comparisons in the single-shot setting as Figure 4 in the main paper.

A.5. More qualitative results

We show more qualitative results of our method showing the effect of increasing the K-shot inputs, and the effect of applying the subject fine-tuning in Figure 5. We observe that we get a noticeable improvement when we increase K from 1 to 4. The visual gain from increasing K further starts to saturate, although quantitative metrics generally keep improving (*e.g.,* Table 2). While increasing K beyond 4 still leads to better visual results in general, we observe that the most improvement focuses on the background and clothes reconstruction, with slight improvements to sharpness and color matching as well. Subject fine-tuning further improves the sharpness and better reconstructs the background details.

A.6. More reenactment results

We first expand Figure 7 of the main paper by showing the same reenactment results but for the single-shot meta-learned setting and the 4-shot inputs in both the meta-learned and subject fine-tuned settings in Figure 6. The results show that even in the single-shot meta-learned setting, our model does a pretty good job extrapolating the input image (source) to challenging poses and expressions, while preserving the source identity. Increasing the input shots to 4 leads to a noticeable visual improvement, and fine-tuning further leads to slight improvements, most notable in the female source

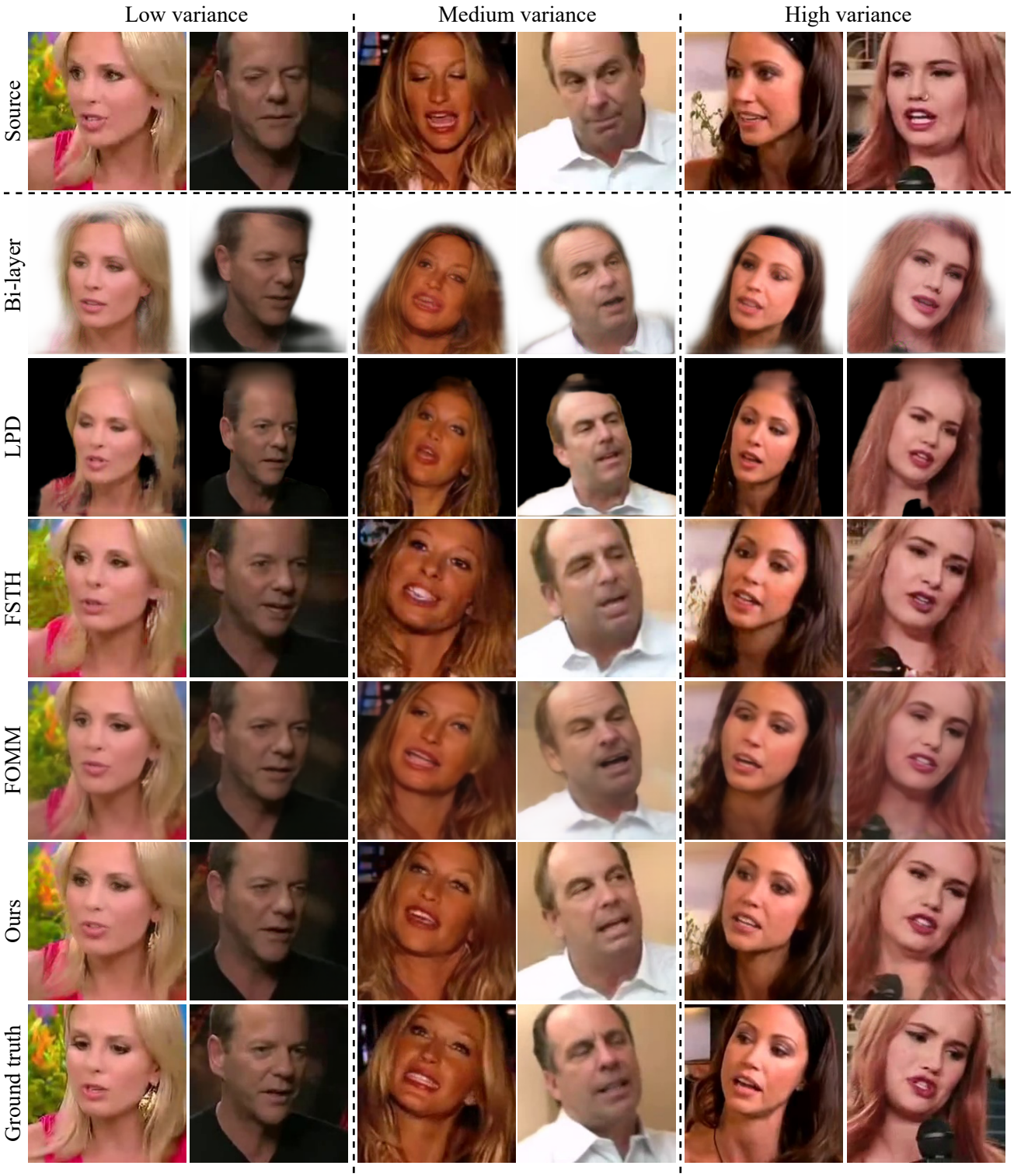

Figure 4: Extending Figure 4 of the main paper, showing more qualitative comparisons in the single-shot setting. We show three sets of examples representing low, medium and high variance between the source and target poses. Our method is more robust to pose variations than the baselines.

Figure 5: Qualitative results of our method showing the gains of increasing the number of K-shot inputs and applying subject fine-tuning.

Figure 6: Expanding Figure 7 of the main paper by showing the same reenactment results in the single and 4-shot settings. Our model extrapolates well to challenging poses and expressions even with a single-shot input (shown in source), while preserving the source identity.

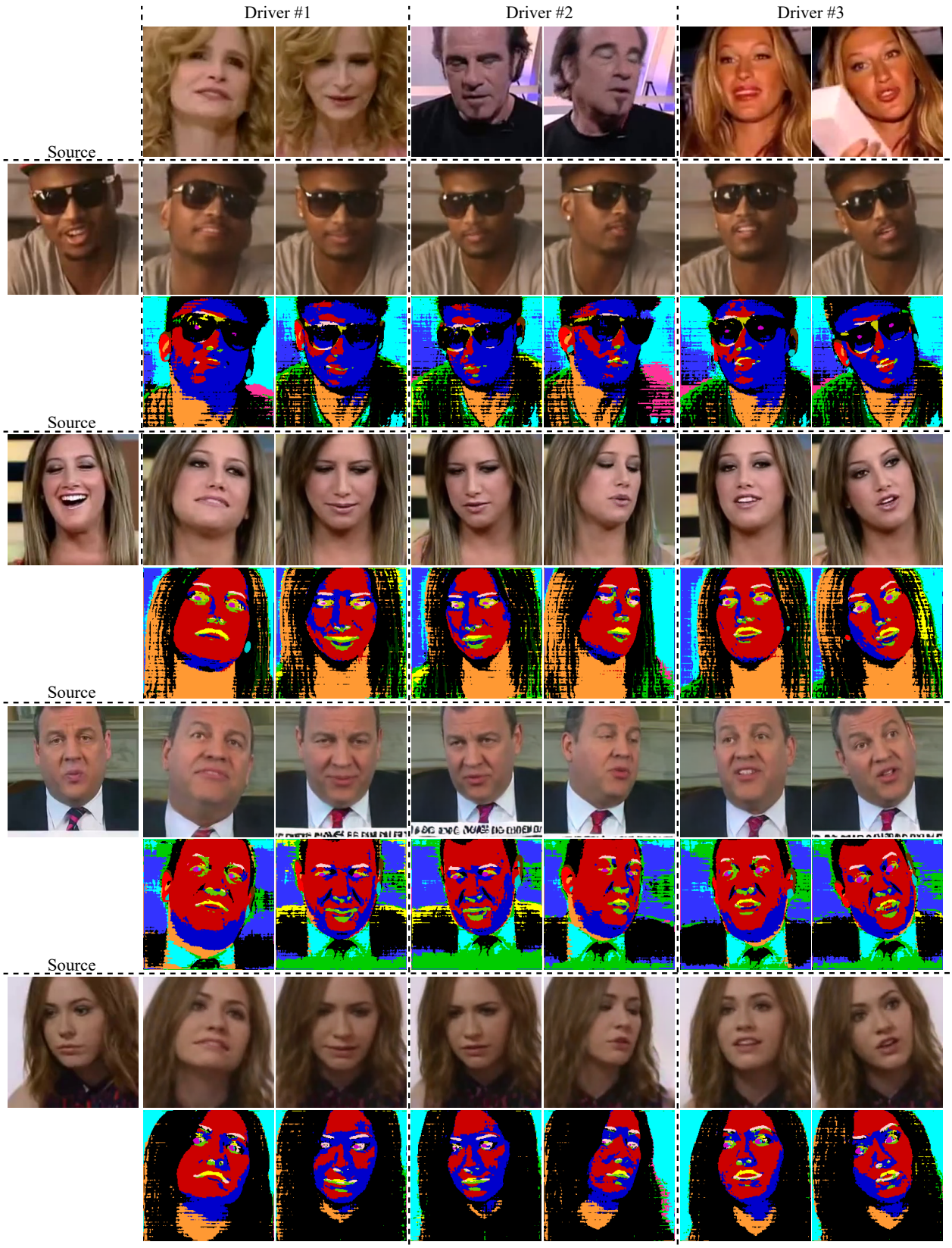

Figure 7: More cross-subject reenactment results with different driving identities. Results are shown for our *meta-learned* model without any fine-tuning, and using 32-shot inputs. We also show the corresponding latent spatial layout maps.

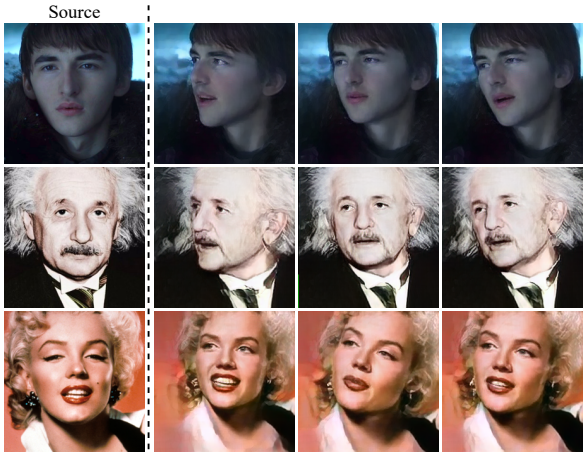

Figure 8: Qualitative results on subjects not belonging to VoxCeleb2.

(middle example). These results show that our method does not require many frames to produce realistic and identity preserving results. For video comparison with the baselines, please refer to the supplementary video.

We also extend Figure 7 of the main paper by showing more reenactment results in Figure 7. We also show the predicted spatial layouts corresponding to the outputs. The predicted spatial layouts may be less interpretable than traditional semantic segmentations, but they seem to encode more information and capture accurate details about the face shape.

Additionally, we perform out-of-domain reenactment using source subjects not present in the VoxCeleb2 dataset. Some qualitative results are shown in Figure 8. Our approach can synthesize realistic novel views given only a single-shot input, although in some cases it shows a bit of an identity gap.

A.7. Limitations and failure cases

Here we discuss some of the limitations of our approach.

Temporal consistency. Similar to previous direct synthesis approaches [3, 14, 15], our training does not enforce temporal consistency. Therefore the output videos often contain some flickering. Considering the temporal aspect during training (*e.g.,* similar to [16]) could mitigate this problem, but on the expense of higher training cost.

Failure modes for cross-subject reenactment. We observe that most of the failure cases in cross-subject reenactment are caused by either source subjects with complex backgrounds, or using male drivers to animate female sources (*e.g.,* Figure 9). Since complex backgrounds could lead to artifacts in our results, then this signifies that the background information is entangled with the face and identity information. Learning a better disentangled representation could improve this problem. On the other hand, the trouble faced with male-to-female reenactment implies that our approach still has some sensitivity to the driver landmarks. While our approach reduces this sensitivity significantly compared to previous baselines, there is still room for improvement.

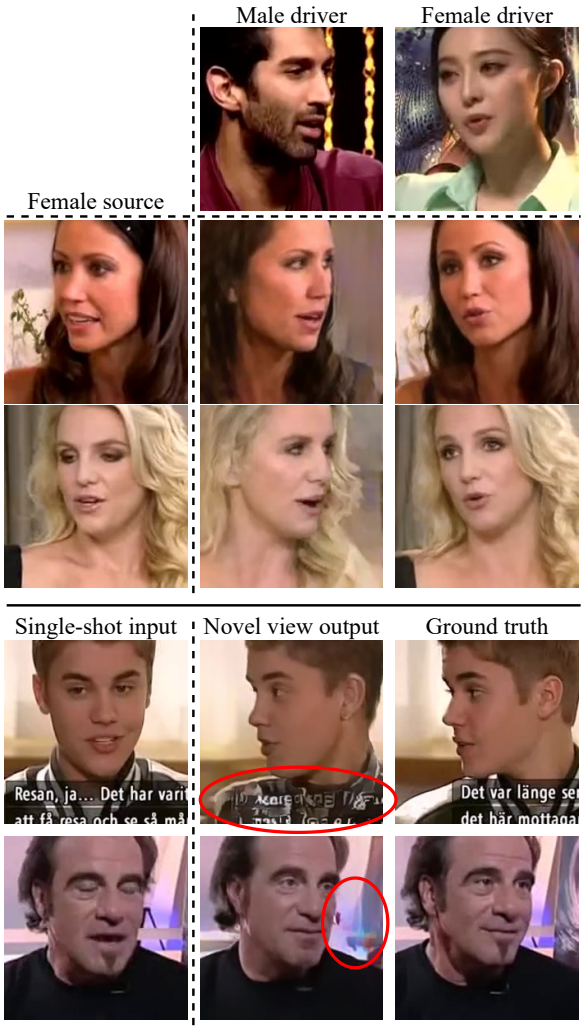

Figure 9: Example limitations. Top: male-to-female reenactment sometimes causes low identity preservation and other visible artifacts. Bottom: our approach cannot faithfully reconstruct the background details.

Background reconstruction Direct synthesis approaches, including our method, synthesize the target frame from a compressed latent code. This compressed bottleneck leads to the loss of some information, especially for the background details. Figure 9 shows some examples in the single-shot setting. Our method cannot transfer static parts (*e.g.,* the closed captions or the background) from the source image to the synthesized view. Borrowing elements from the warping-based approaches is one direction to better reconstruct static details.

Dataset-induced limitations. The VoxCeleb2 [1] has low resolution videos and is processed to perform zoomed-in center crops that often cut off the top of the head. Dataset

biases are inevitably inherited by the trained models. Therefore, generating output for out-of-domain inputs requires pre-processing the inputs to have similar properties to the VoxCeleb2 dataset.

A.8. Ethical concerns

While the task of synthesizing realistic *talking heads* has a wide range of applications, it also raises ethical concerns regarding potential misuses of this technology. A prime example of this is the growing misuse of DeepFakes [17, 18]. Several state-of-the-art methods can easily swap identities, expressions as well as face attributes and generate photo-realistic samples. Additionally, with the increase in the ease of access to face reenactment models, more and more people can misuse such models through widely available applications. Thus it is important at the same time to have the ability to detect fake content. In this direction recent works like [19, 20, 21] have tried to solve the problem of detecting real *vs*. fake images. Especially interesting is the work by Wang *et al*. [20] which shows that models for fake image detection can be made to generalize well to unseen scenarios. While this is a temporary respite, it is important to continue research in the field of fake image detection to keep on par with the ever improving field of image synthesis, as not only do models improve, but also the ease of access to such models grows rapidly.

References

[1] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*, 2018. 1, 8

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Int. Conf. Comput. Vis.*, pages 1021–1030, 2017. 1

[3] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9459–9468, 2019. 1, 2, 3, 8

[4] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *Int. Conf. Learn. Represent.*, 2019. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[7] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Int. Conf. Comput. Vis.*, pages 1501–1510, 2017. 1

[8] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020. 1

[10] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 1

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Int. Conf. Learn. Represent.*, 2018. 1

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Int. Conf. Learn. Represent.*, 2015. 2

[13] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, December 2019. 2

[14] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13786–13795, 2020. 2, 8

[15] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *Eur. Conf. Comput. Vis.*, August 2020. 8

[16] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *NeurIPS*, 2019. 8

[17] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 9

[18] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021. 9

[19] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389. IEEE, 2018. 9

[20] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 9

[21] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019. 9