Pacific Graphics 2022
N. Umetani, E. Vouga, and C. Wojtan
(Guest Editors)

Volume 41 (2022), Number 7

Learning Dynamic 3D Geometry and Texture for Video Face Swapping (Supplemental Material)

C. Otto[1,2], J. Naruniec[1], L. Helminger[1,2], T. Etterlin[2], G. Mignone[1], P. Chandran[1,2], G. Zoss[1], C. Schroers[1], M. Gross[1,2], P. Gotardo[1], D. Bradley[1], R. Weber[1]

[1]DisneyResearch|Studios, Switzerland
[2]ETH Zürich, Switzerland

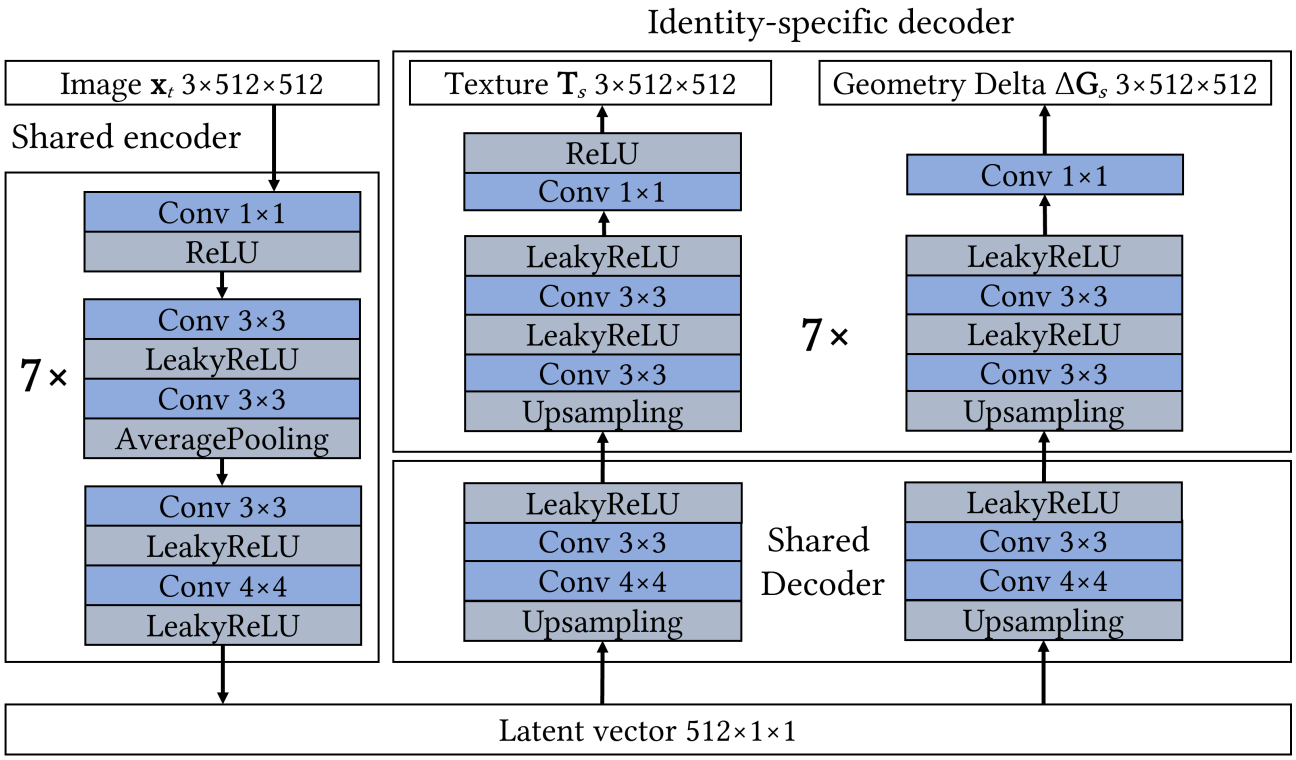

Figure 1: *Details of our encoder and decoder network architecture. We adapt the basic comb model of [NHSW20] with separate texture and geometry decoders per identity to disentangle both domains. The encoder and the first part of the two decoders are shared across both identities, to achieve an accurate expression transfer [NHSW20].*

In this supplemental material we provide insights into the details of our implementation, discuss the effect of illumination on our results, present the findings of a user study, and show further face swaps.

1. Implementation Details

Network Architecture Details Figure 1 shows the details of our autoencoder network. We extend the basic comb model of [NHSW20] by two identity-specific decoders. One for learning the dynamic face texture and one for learning the geometry deltas.

Face model Our method requires initializing the static geometry with a coarse face mesh $\mathbf{S}$ at the start of training. We also require image-specific pose parameters $\mathbf{R}$ and $\mathbf{t}_{2d}$, respectively the rotation and translation, which are applied to the learned geometry before rendering. For both purposes we utilize the 3D Dense Face Alignment (3DDFA) model with clipped parameters as in

[GZL18, ZLLL17, GZY*20]. 3DDFA is based on the 3DMM from Blanz et al. [BV03], which utilizes PCA to construct 3D faces.

The 3D face can be described as $\mathbf{S} = \bar{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}$, where $\bar{\mathbf{S}}$ is the mean shape, $\mathbf{A}_{id}$ is the principal axis, and α_{id} the shape parameter of the identity's neutral expression from the Basel Face Model [PKA*09]. The principal axis $\mathbf{A}_{exp}$ is trained with the displacement between expression and neutral face scans from FaceWarehouse [CWZ*14], where α_{exp} is the expression parameter. In total 3DDFA provides the model parameters $\mathbf{w} = [f, \mathbf{R}, \mathbf{t}_{2d}, \alpha_{id}, \alpha_{exp}]^\top$, which can be used to project the 3D face to the 2D image plane: $V(\mathbf{w}) = f * \mathbf{Pr} * \mathbf{R} * \mathbf{S} + \mathbf{t}_{2d}$, where $\mathbf{Pr}$ is the orthographic projection matrix and f the scale factor [ZLLL17]. To fit the face model to an image, we downscale our normalized image to 120×120, as this is the input size that 3DDFA requires. We apply parameter averaging to improve the face model fit as well as the temporal stability. To achieve better temporal stability, we shift the bounding box derived from the landmarks in the image plane by $\beta\gamma$ with strength $\beta = 0.05$, where γ is the width of the bounding box, and perform $n = 9$ face model fits per image [NHSW20, Ett20]. The average parameters across all n fits produce the final face parameters, which are saved for training. We also adjust the 3DDFA model crop for face swapping purposes by removing the ears and neck. We then close the mouth and nostrils in order to learn a texture for these face parts. We take the UV coordinates for the Basel Face Model from Bas et al. [BHS*17] and adjust them to our cropped mesh via an affine transformation. Next, similar to [Ett20], we scale the mouth region by a non-uniform multivariate Gaussian to learn a more detailed representation of the teeth and lips.

2. Effect of Illumination

In the main document, we explained that the dataset from Naruniec et. al. [NHSW20] contains sequences from 8 identities under 3 different lighting conditions, which we use to train networks to swap between pairs of identities. Importantly, not all identities were captured under all lighting conditions, leading to two different swap scenarios - one where the same illumination conditions for both the source and target of a particular swap were seen by the network during training, and a second more challenging scenario where the source identity was not seen under the target illumination of a particular swap during training.

Our results indicate that swaps look visually more realistic when the illumination condition of the target video was also available for the source identity at training time, as compared to the second scenario where the source identity lit by the target illumination was unavailable. Some examples are shown in Figure 2. We believe this limitation comes from the fact that illumination is not explicitly modeled by our method, and thus the texture prediction network must predict both albedo as well as shading for each frame. Explicitly incorporating illumination and allowing the texture network to predict pure albedo could alleviate this issue, and would be an interesting avenue for future work.

3. Perceptual User Study

We conducted a perceptual user study with 59 participants. The goal of the study was to visualize dynamic (video) swap results generated from our method along with four state of the art methods [NHSW20, PGC*20, NKH19, CCNG20], and ask the user several questions about the swap quality.

Each participant saw four video clips that show each of the methods performing face swaps on the same sequence. To create the four test clips fairly, we randomly selected four 10-second subsequences from all our available data. Figure 3 shows a screenshot from each of the four clips in the user study. Participants saw the methods in each video in random order. For evaluation, we asked for the method that best retains the identity of the source person (identity), the one that keeps the target lighting, pose and expression of the target the best (attributes) and the one that displays the most realistic video swap (realism). The questions are similar to those asked in [LBY*19] but applied to the video context. In total the participants had to answer 18 questions, and were asked to provide the top three votes per question, in order to rank the top three performing methods.

The results of the user study are summarized in Table 1, which sums the total number of votes per category across all subsequences. The summary table indicates that no single method outperforms the others in all categories, and also that our 3D-based approach performs comparably to existing 2D methods.

Method	Identity	Attributes	Realism
Naruniec et al. [NHSW20]	181	458	208
DFL [PGC*20]	196	364	178
FSGAN [NKH19]	50	165	29
SimSwap [CCNG20]	103	464	162
Ours	178	319	131

Table 1: *The total number of user votes per category across all subsequences.*

To further analyze the results we refer back to Figure 3, where we see that three out of the four randomly picked sequences contain source identity and target performance in different lighting conditions (sequences 1, 2 and 4). Only one sequence (sequence 3) captures them in similar lighting conditions. As discussed in the previous section, our method performs visually better in the latter case. This is verified by the user study, where our method ranks first in terms of user votes for both identity preservation and realism on this illumination-consistent sequence, indicating that our method has potential to perform very well when lighting conditions are the same. On the other three videos, Naruniec et al. [NHSW20] followed by DeepFaceLab [PGC*20] rank best, accumulating the most user votes across all questions. While the magnitude of our study is perhaps not large enough to make definitive conclusions (with only 4 test videos), we believe our method shows large improvements over previous 3D-based face swapping methods such as [NMT*18], and is starting to close in on the performance gap between 2D and 3D-based face swapping methods. While related work such as [LBY*19] performs a user study on individual frames, our study is performed on dynamic video swaps, in the hopes of achieving a more meaningful comparison.

Figure 2: *Swapping in different lighting condition makes it more difficult for our method to create photorealistic swaps (left), because a lot of the light is baked into our texture. In contrast, when the lighting conditions of source identity and target performance are the same our method performs very well (right).*

4. Additional Face Swaps

We show further face swaps created by our method in Figure 4.

References

[BHS*17] BAS A., HUBER P., SMITH W. A. P., AWAIS M., KITTLER J.: 3d morphable models as spatial transformer networks. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (2017), pp. 895–903. 2

[BV03] BLANZ V., VETTER T.: Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell. 25*, 9 (2003), 1063–1074. 2

[CCNG20] CHEN R., CHEN X., NI B., GE Y.: *SimSwap: An Efficient Framework For High Fidelity Face Swapping*. Association for Computing Machinery, New York, NY, USA, 2020, p. 2003–2011. 2

[CWZ*14] CAO C., WENG Y., ZHOU S., TONG Y., ZHOU K.: Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics 20* (2014), 413–425. 2

[Ett20] ETTERLIN T.: Leveraging 3d geometry for neural face swapping. *Masterthesis* (2020). 2

[GZL18] GUO J., ZHU X., LEI Z.: 3ddfa. 2

[GZY*20] GUO J., ZHU X., YANG Y., YANG F., LEI Z., LI S. Z.: Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2020). 2

[LBY*19] LI L., BAO J., YANG H., CHEN D., WEN F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457* (2019). 2

[NHSW20] NARUNIEC J., HELMINGER L., SCHROERS C., WEBER R.: High-resolution neural face swapping for visual effects. *Computer Graphics Forum 39*, 4 (2020), 173–184. 1, 2

[NKH19] NIRKIN Y., KELLER Y., HASSNER T.: FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 7184–7193. 2

[NMT*18] NIRKIN Y., MASI I., TRAN A. T., HASSNER T., MEDIONI G.: On face segmentation, face swapping, and face perception. In *IEEE Conference on Automatic Face and Gesture Recognition* (2018). 2

[PGC*20] PEROV I., GAO D., CHERVONIY N., LIU K., MARANGONDA S., UMÉ C., DPFKS M., FACENHEIM C. S., RP L., JIANG J., ZHANG S., WU P., ZHOU B., ZHANG W.: Deepfacelab: A simple, flexible and extensible face swapping framework. *CoRR abs/2005.05535* (2020). 2

[PKA*09] PAYSAN P., KNOTHE R., AMBERG B., ROMDHANI S., VETTER T.: A 3d face model for pose and illumination invariant face recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments* (Genova, Italy, 2009), IEEE. 2

[ZLLL17] ZHU X., LIU X., LEI Z., LI S. Z.: Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence* (2017). 2

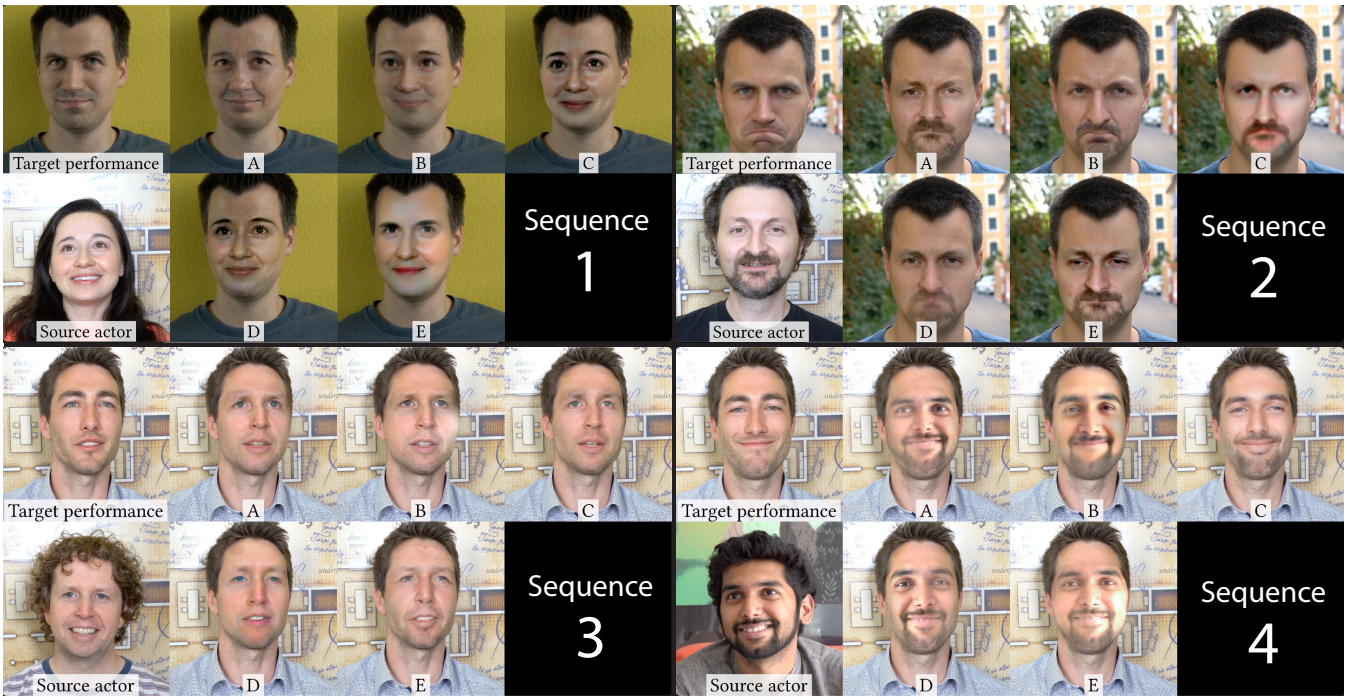

Figure 3: *Four screenshots from the four sequences in the user study, showing video results from all methods at the same time.*

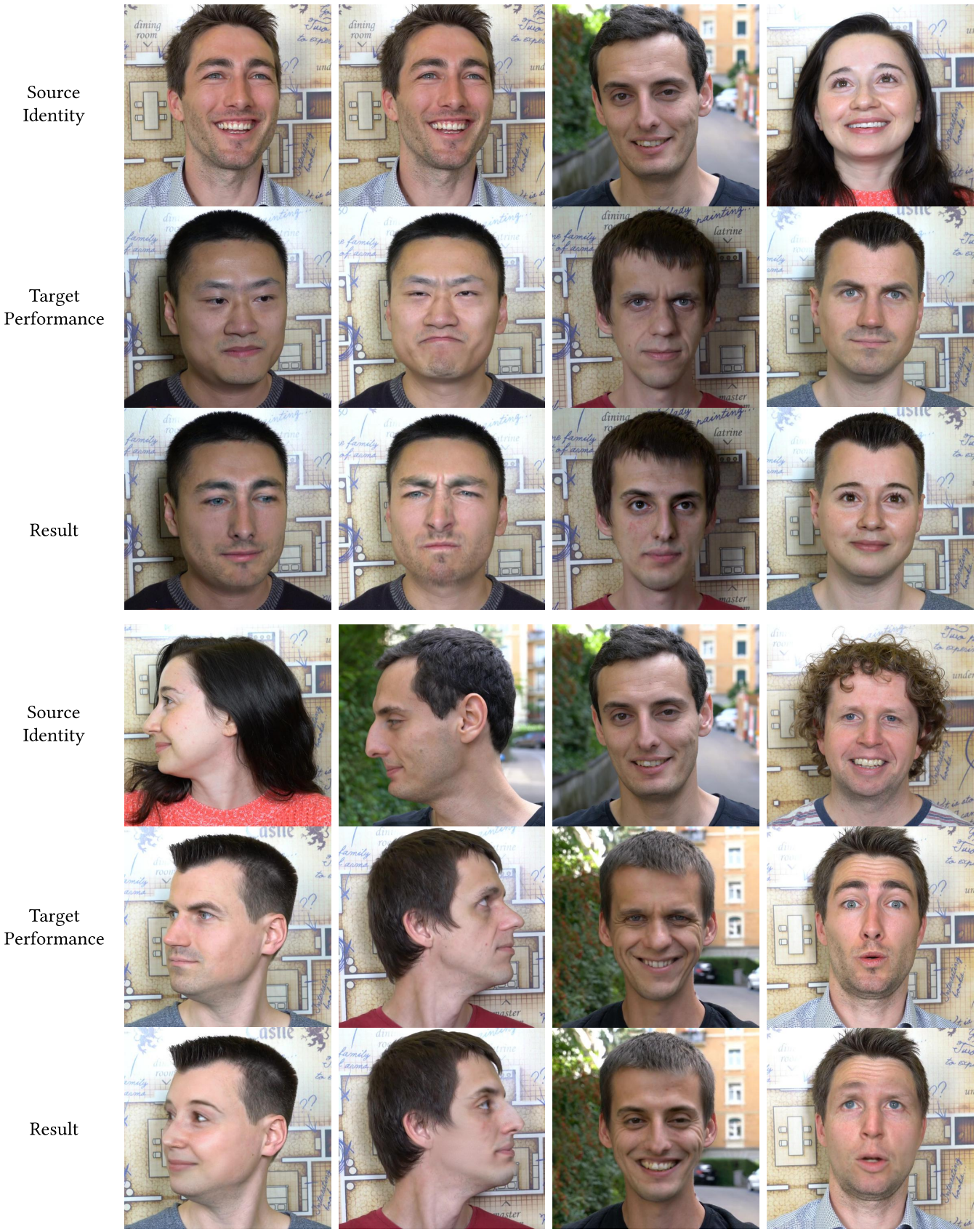

Figure 4: *Additional face swaps created by our method.*