Neural Voice Puppetry: Audio-driven Facial Reenactment

–Supplemental Document–

Justus Thies[1], Mohamed Elgharib[2], Ayush Tewari[2], Christian Theobalt[2], and Matthias Nießner[1]

[1] Technical University of Munich
[2] Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract. In this supplemental document, we give additional information to our method *Neural Voice Puppetry*. Specifically, we detail on the used network architectures for the audio-expression network and the rendering network, as well as the training. We report the statistics of our user study that evaluated visual quality and audio-visual sync, and provide additional comparisons to state-of-the-art methods. The supplemental material is concluded with a section about ethical considerations.

1 Network Architectures

Audio2ExpressionNet: A core component of *Neural Voice Puppetry* is the estimation of facial expressions based on audio. To retrieve temporal coherent estimations, we employed a process with two stages. In the first stage, we estimate per frame expressions based on DeepSpeech features. The output of this network is an audio-expression vector of length 32. This audio-expression is temporally noisy and is filtered using an expression aware filtering network which can be trained in conjunction with the per frame expression estimation network. The temporal filtering mechanism is also depicted in the main paper. The underlying network that predicts the filter weights gets as input $T = 8$ per-frame predicted audio expressions. We apply 5 1D-convolutional filters with kernel size 3 that reduce the feature space successively from 8×32 over 8×16, 8×8, 8×4, 8×2 to 8×1. Each of these convolutions has a bias and is followed by a leaky ReLU activation (negative slope of 0.02). The output of the convolutional network is input to a fully connected layer with bias that maps the 1×8 input to the 8 filter weights that are normalized using a softmax function. To train the network we apply a vertex-based loss as described in the main paper. The vertices that refer to the mouth region are weighted with a $10\times$ higher loss. We use the mask that is depicted in Fig. 1. For generalization we used a dataset composed of commentators from the German public TV (e.g., https://www.tagesschau.de/multimedia/video/video-587039.html). In total the dataset contained 116 videos.

Fig. 1. Mask.

Rendering network: In Fig. 2, we show an overview of our neural rendering approach. Based on the expression predictions, that drive a person-specific 3D face model, we render a neural texture to the image space of the target video. A first network is used to convert the neural descriptors sampled from the neural texture to RGB color values. A second network embeds this image into the target video frame. We erode the target image around the synthetic image, to remove motions of the target actor like chin movements. Using this eroded target image as background and the output of the first network, the second network outputs the final image.

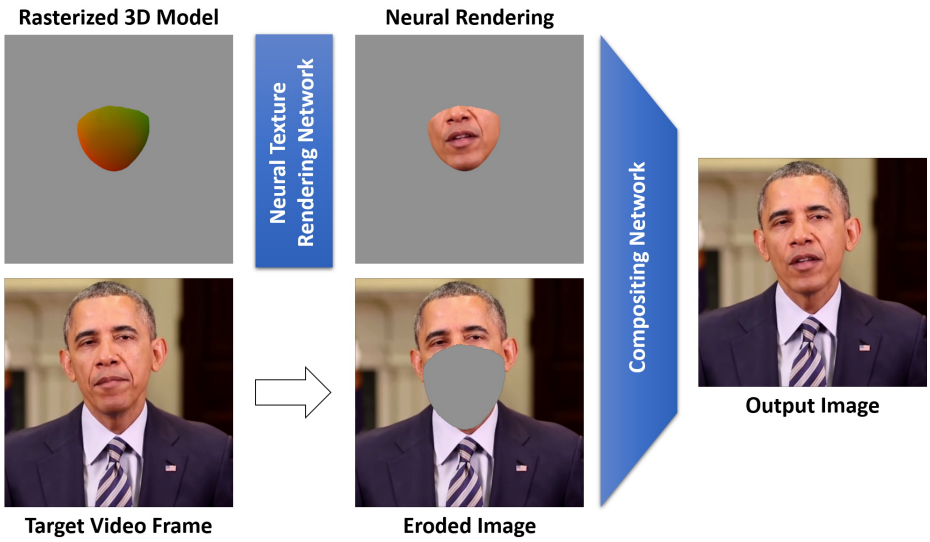

Fig. 2. Our neural rendering approach consists of a deferred neural renderer and an inpainting network that blends the modified face interior into the target image.

Both networks have the same structure, only the input dimensions are different. The first network gets an image with 16 feature channels as input (dimension of the neural descriptors that are sampled from a neural texture with dimensions $256 \times 256 \times 16$), while the second network composites the background and the output of the first network, resulting in an 6 channel input. The networks are implemented in the Pix2Pix framework [6]. Instead of a classical U-Net with strided convolutions, we build on dilated convolutions. Specifically, we replace the strided convolutions in a U-Net of depth 5. Instead of transposed convolutions, we use standard convolutions, since we do not downsample the image and always keep the same image dimensions. Note that we also keep the skip connections of the classical U-Net. The number of features per layer is 32 in our experiments, resulting in networks with $\sim 2.35mio$ parameters (which is low in comparison to the network in Deferred Neural Rendering [13] with $\sim 16mio$ pa-

rameters). We employ the structure that is depicted in Fig. 3. Each convolution layer has a kernel size of 3×3 and is followed by a leaky ReLU with negative slop of 0.2. All layers have stride 1 which means that all layers intermediate feature maps have the same spatial size as the input (512×512). The first convolutional layer maps to a feature space of dimension 32 and has a dilation of 1. With increasing layer depth the feature space dimension as well as the dilation increases by a factor of 2. After layer depth 5, we use standard convolutions.

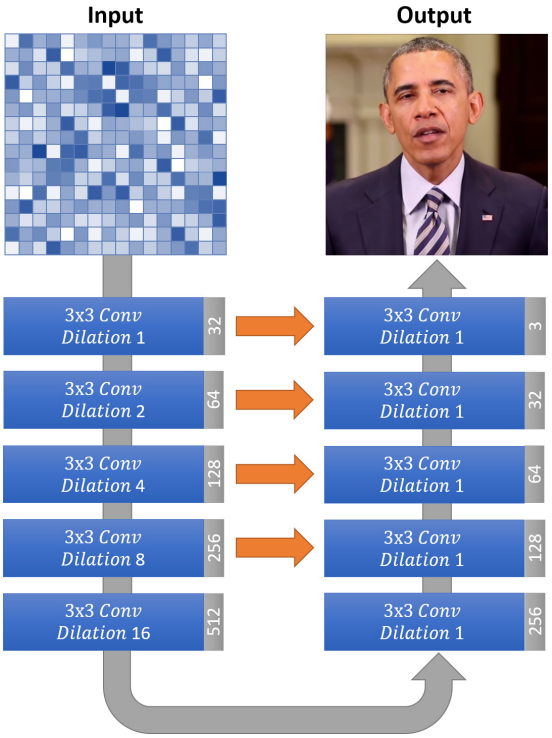

Fig. 3. We use a modified U-Net architecture that uses dilated convolutions instead of strided convolutions. Transposed convolutions are replaced by std. convolutions.

Training Our pipeline is implemented in PyTorch using the Adam [10] optimizer with default settings ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \cdot e^{-8}$), a learning rate of 0.0001 and Xavier initialization. The *Audio2ExpressionNet* is trained for 50 epochs (resulting in a training time of $\sim$ 28 hours on an Nvidia 1080Ti) with a learning rate decay for the last 30 epochs and a batch size of 16. The rendering networks are trained for 50 epochs for each target person individually with a batch size of 1 ($\sim$ 30 hours training time, $\sim$ 5 hours in case of strided convolutions).

2 User Study

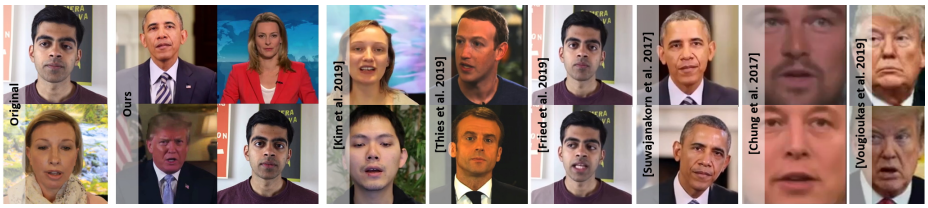

Fig. 4. Our user study contained 24 videos from different state-of-the-art methods, including 3 original videos. Here we show some frames of the videos.

In this section, we present the statistics of our user study. Fig. 5 shows a collection of videos that we used for the user study. The clips are from the official videos of the corresponding methods and are similar to the clips that we show in our supplemental video. Fig. 5 shows the average answers of our questions, including the variance.

In the user study we asked the following questions:

- How would you rate the audio-visual alignment (lip sync) in this video?
- How would you rate the visual quality of the video?

With the answer possibilities "very good","good","Neither good nor bad","bad", "very bad".

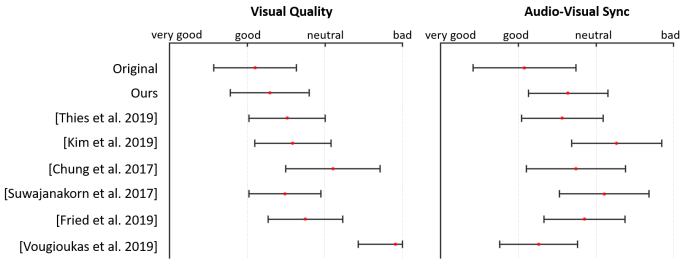

Fig. 5. Statistics of our user study including the mean and the variance with respect to the specific methods and question about visual quality and audio-visual sync (lip sync).

3 Additional Comparisons to State-of-the-Art

Image-based & Audio-driven Facial Animation: In addition to the results in the main paper, we also compare to Chen et al. [1] and Vougioukas et al. [14]. For both methods, we use publicly available pretrained models [3][4]. We compare on a sequence of Obama in a self-reenactment scenario (to provide ground truth images). As can be seen in Fig. 6, our method surpasses the visual image quality of these methods and generates full frame images (in contrast to normalized facial images). Since the image-based methods are operating in a normalized space, we cannot provide a fair quantitative evaluation w.r.t. the ground truth images. To compute PSNR and landmark errors, we would need to transform the ground truth images to the normalized space which leads to errors since the head is moving and we can not assume perfect tracking of the face bounding box. Especially, rotations of the head are also not handled by the image-based methods which would dominate the error. Our PSNR and landmark errors for the self-reenactment sequence of Obama are listed in the main paper.

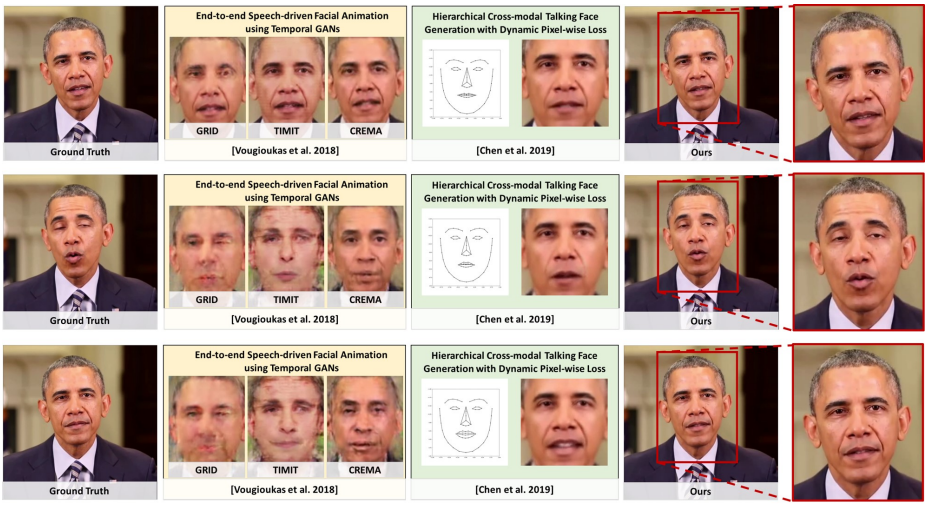

Fig. 6. Comparison to Chen et al. [1] and Vougioukas et al. [14] on a sequence of Obama (self-reenactment). From top to bottom we show the frames 0, 70 and 305. Note that we list 3 results for the method of Vougioukas et al. which are based on different training datasets (GRID, TIMIT, CREMA). The respective sequence is part of the supplemental video.

[3] https://github.com/lelechen63/ATVGnet
[4] https://github.com/DinoMan/speech-driven-animation

Model-based Audio-driven Facial Animation: In our supplemental video, we show multiple comparisons to Voca [3]. Fig. 7 shows an image of a legacy Winston Churchill sequence. In contrast to Voca, our aim is to generate photo-realistic output videos that are in sync with the audio. Voca focuses on the 3D geometry requiring a 4D training corpus, while our approach uses a 3D proxy only as an intermediate step and works on videos from the Internet. Our 3D proxy is based on a generic face model and, thus, has not the details as a person-specific modelled mesh. Nevertheless, using our neural rendering approach, we are able to generate photo-realistic results.

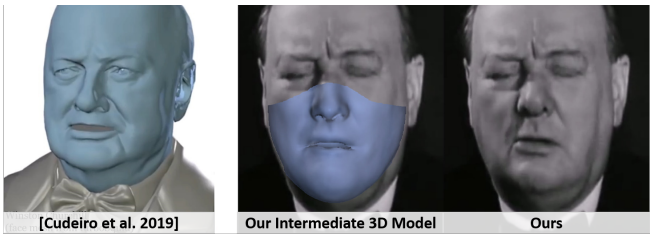

Fig. 7. Qualitative comparison of our method to Voca [3]. It is a representative image for a talking sequence of Winston Churchill.

Model-based Video-driven Dubbing & Facial Reenactment: State-of-the-art video dubbing is based on video-driven facial reenactment [5,12,9,13,8]. In contrast, our method is only relying on the voice of the dubber. The 'Deferred Neural Rendering' [13] is a generic neural rendering approach, but the authors also show the usage in the scenario of facial reenactment. It builds upon the Face2Face [12] pipeline and directly transfers the deformations from the source to the target actor. Thus, tracking errors that occur in the source video (e.g., due to occlusions or fast motions) are transferred to the target video. In a dubbing scenario, the goal is to keep the talking style of the target actor which is not the case for [5,12,9,13]. To compensate the influence of the source actor talking style, Kim et al. [8] proposed a method to map from the source style to the target actor style. Our approach directly operates in the target actor expression space, thus, no mapping is needed (we also do not capture the source actor style). This enables us to also work on strong expressions, as shown in Fig. 8.

Text-driven Video Synthesis: Fried et al. presented 'Text-based Editing of Talking-head Video' [4] which provides a video editing tool that is based on the transcript of the video. The method reassembles captured expression snippets from the target video, requiring blending heuristics. To achieve their results they rely on more than one hour of training data. We show a qualitative comparison to this method in the supplemental video. Our method only uses the synthetic

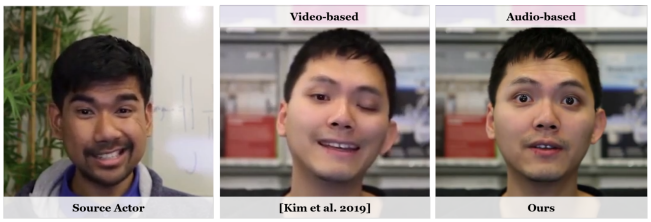

Fig. 8. Visual dubbing fails to map strong expressions from the source to plausible expressions of the target actor.

audio sequence as input, while the method of Fried et al. uses both the transcript and the audio. Note that our method generates the entire video, while the text-based editing method only synthesizes the frames of the new three words.

4 Ethical Considerations

In conjunction with person specific audio generators like Jia et al. [7], a pipeline can be established that creates video-realistic (temporal voice- and photo-realistic) content of a person. This is perfect for creative people in movie and content production, to edit and create new videos. On the other hand, it can be misused. To this end, the field of digital media forensics is getting more attention. Recent publications [11] show that humans have a hard time in detecting fakes, especially, in the case of compressed video content. Learned detectors are showing promising results, but are lacking generalizeability to other manipulation methods that are not in the training corpus. Few-shot learning methods like ForensicTransfer [2] try to solve this issue. As part of our responsibility, we are happy to share generated videos of our method with the forensics community. Nevertheless, our approach enables several practical use-cases, ranging from movie-dubbing to text-driven photo-realistic video avatars. We hope that our work is a stepping stone in the direction of audio-based reenactment and is inspiring more follow-up projects in this field.

References

1. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7832–7841 (2019) 5
2. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv (2018) 7
3. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.: Capture, learning, and synthesis of 3D speaking styles. Computer Vision and Pattern Recognition (CVPR) (2019) 6

4. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M.: Text-based editing of talking-head video. ACM Trans. on Graph. (Proceedings of SIGGRAPH) **38**(4), 68:1–68:14 (Jul 2019) 6
5. Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., Theobalt, C.: VDub - modifying face video of actors for plausible visual alignment to a dubbed audio track. In: Computer Graphics Forum (Proceedings of EUROGRAPHICS) (2015) 6
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arxiv (2016) 2
7. Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I.L., Wu, Y.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: International Conference on Neural Information Processing Systems (NIPS). pp. 4485–4495 (2018) 7
8. Kim, H., Elgharib, M., Zollhöfer, M., Seidel, H.P., Beeler, T., Richardt, C., Theobalt, C.: Neural style-preserving visual dubbing. ACM Trans. on Graph. (SIGGRAPH Asia) (2019) 6
9. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Trans. on Graph. (Proceedings of SIGGRAPH) **37**(4), 163:1–163:14 (Jul 2018) 6
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980 3
11. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. arXiv (2019) 7
12. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In: CVPR (2016) 6
13. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Trans. on Graph. (Proceedings of SIGGRAPH) (2019) 2, 6
14. Vougioukas, K., Petridis, S., Pantic, M.: End-to-end speech-driven facial animation with temporal gans. In: BMVC (2018) 5