Neural Style-Preserving Visual Dubbing

HYEONGWOO KIM, Max Planck Institute for Informatics
MOHAMED ELGHARIB, Max Planck Institute for Informatics
MICHAEL ZOLLHÖFER, Stanford University
HANS-PETER SEIDEL, Max Planck Institute for Informatics
THABO BEELER, DisneyResearch|Studios
CHRISTIAN RICHARDT, University of Bath
CHRISTIAN THEOBALT, Max Planck Institute for Informatics

Fig. 1. Our visual dubbing method enables style-preserving lip synchronization by translating the source actor's facial expressions to a target actor's idiosyncratic style. Current dubbing techniques perform direct expression transfer from source to target actors. This reproduces the facial expressions of the source actor and leads to the loss of the style and idiosyncrasies of the target actor.

Dubbing is a technique for translating video content from one language to another. However, state-of-the-art visual dubbing techniques directly copy facial expressions from source to target actors without considering identity-specific idiosyncrasies such as a unique type of smile. We present a style-preserving visual dubbing approach from single video inputs, which maintains the signature style of target actors when modifying facial expressions, including mouth motions, to match foreign languages. At the heart of our approach is the concept of motion style, in particular for facial expressions, i.e., the person-specific expression change that is yet another essential factor beyond visual accuracy in face editing applications. Our method is based on a recurrent generative adversarial network that captures the spatiotemporal co-activation of facial expressions, and enables generating and modifying the facial expressions of the target actor while preserving their style. We train our model with unsynchronized source and target videos in an unsupervised manner using cycle-consistency and mouth expression losses, and synthesize photorealistic video frames using a layered neural face renderer. Our approach generates temporally coherent results, and handles dynamic backgrounds. Our results show that our dubbing approach maintains the idiosyncratic style of the target actor better than previous approaches, even for widely differing source and target actors.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Neural networks**; *Appearance and texture representations*; *Animation*; *Rendering*.

Additional Key Words and Phrases: Visual Dubbing, Motion Style Transfer, Generative Adversarial Networks, Recurrent Neural Networks

ACM Reference Format:
Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. 2019. Neural Style-Preserving Visual Dubbing. *ACM Trans. Graph.* 38, 6, Article 178 (November 2019), 13 pages. https://doi.org/10.1145/3355089.3356500

Authors' addresses: Hyeongwoo Kim, Max Planck Institute for Informatics, hyeongwoo.kim@mpi-inf.mpg.de; Mohamed Elgharib, Max Planck Institute for Informatics, elgharib@mpi-inf.mpg.de; Michael Zollhöfer, Stanford University, zollhoefer@cs.stanford.edu; Hans-Peter Seidel, Max Planck Institute for Informatics, hpseidel@mpi-sb.mpg.de; Thabo Beeler, DisneyResearch|Studios, thabo.beeler@disneyresearch.com; Christian Richardt, University of Bath, christian@richardt.name; Christian Theobalt, Max Planck Institute for Informatics, theobalt@mpi-inf.mpg.de.

0730-0301/2019/11-ART178 $15.00
https://doi.org/10.1145/3355089.3356500

1 INTRODUCTION

Localization of media content, such as feature films or television series, has nowadays become a necessity since the target audience is oftentimes not familiar with the original spoken language. The process to replace the original dialog with a different language, typically spoken by a different voice actor, is known as *dubbing*. The challenge that arises in traditional audio-only dubbing is that the audio and visual signals do not match anymore. This is not only distracting but can significantly reduce understanding, since up to one third of speech information is captured from the visual signal in the presence of noise [Le Goff et al. 1994]; and this is obviously

aggravated for hearing-impaired viewers who rely on lip reading [Owens and Blazek 1985]. Hence research (see Section 2), and more recently industry (e.g., https://synthesia.io/), has started to address the problem of *visual* dubbing, where the visual content is adjusted to match the new audio channel.

Every person speaks in a unique way, both in terms of expressions as well as their timing. In particular, for actors, politicians and other prominent people, their idiosyncrasies and demeanor are part of their 'brand' and it is of utmost importance to preserve their style when dubbing. However, so far the field has entirely ignored style when dubbing. This causes uncanny results, in particular for well-known actors – just imagine, for example, Robert DeNiro performing with the idiosyncrasies of Sylvester Stallone.

In this work, we propose the first method for visual dubbing that is able to preserve the style of the target actor, allowing to faithfully preserve a person's identity and performance when dubbing to new languages. We achieve this by learning to automatically translate a source performance to a target performance, requiring only unpaired videos of the two actors. We learn how to perform the retargeting in parameter space using a cycle-consistency loss, and utilize a long short-term memory (LSTM) architecture to provide a solution that is temporally coherent. We first convert the videos into our parametric representation using a multilinear face model and finally convert back to the video domain using a novel layer-based neural face renderer, which is capable of handling dynamic backgrounds. More specifically, the novelties presented in this paper include:

- the first approach for visual dubbing that preserves the style of an original actor while transferring the lip movements according to a dubbing actor,
- a novel target-style preserving facial expression translation network that we train in an unsupervised fashion using a temporal cycle-consistency loss and a mouth expression loss, and
- a layer-based approach for neural face rendering that can handle dynamic video backgrounds.

2 RELATED WORK

Traditional dubbing pipelines seek to optimize the alignment between the dubbed audio and the mouth movements of the original target actor. The source actor is recorded in a studio while reading out the dubbing script. The script is translated in a way to maintain the overall semantics of the original script while trying to match the salient mouth movements of the target, such as the bilabial consonants /b/, /m/ and /p/. The dubbing actor reads out the script and attempts to be in pace with the target's voice as much as possible. Finally, the dubbed audio is manually edited to further improve its alignment with the target actor's mouth region. While commercial dubbing pipelines require professional dubbing actors and tedious manual editing, several techniques have been proposed to reduce the complexity of this process. These techniques go back as far as the work of Brand [1999] for voice puppetry. The vast majority of dubbing-related techniques, however, can be divided into audio-based and visual-based approaches.

2.1 Audio-Based Dubbing Techniques

Audio-based dubbing techniques learn to associate the input audio stream of a driving source voice with the visual facial cues of a target actor. This is challenging as there exists no one-to-one mapping between phonemes and visemes, i.e., the same sentence can be said with different expressions. Motion-captured data are commonly used to represent facial visual cues, even in early work [Deng and Neumann 2006; Kshirsagar and Magnenat-Thalmann 2003; Ma et al. 2006; Taylor et al. 2012]. Recently, the rise of deep learning has led to noticeable improvements in audio-based dubbing techniques. Karras et al. [2017] map the raw audio waveform to the 3D coordinates of a face mesh. A trainable parameter is defined to capture emotions. During inference, this parameter is modified to simulate different emotions. Taylor et al. [2017] use a sliding-window deep neural network to learn a mapping from audio phoneme sequences to active appearance model (AAM) parameters. The learned AAM parameters can be retargeted to different face rigs. Pham et al. [2017] and Cha et al. [2018] proposed a neural network approach to learn the mapping to facial expression blendshapes. These approaches do not seek to generate photorealistic videos, and rather focus on controlling facial meshes or rigs for computer animation, or producing cartoon-looking characters. Hence, they cannot be used for dubbing of more general visual content such as movies and TV shows.

The generation of photorealistic images or videos for audio dubbing has only seen little work to date. Chung et al. [2017] presented a technique that animates the mouth of a still image in a way that follows an audio speech. Their approach builds on a joint embedding of the face and audio to synthesize the talking head. Vougioukas et al. [2018] introduce a similar method with a temporal generative model to produce more coherent speech-driven facial animation over time. The end result is an animation of a still image and not yet a natural video. Suwajanakorn et al. [2017] presented an audio-based dubbing technique with high-quality video animation as the end result. A recurrent neural network is trained on 17 hours of President Obama's speeches to learn the mouth shape from the audio. Their approach assumes the source and target have the same identity, and requires many hours of training data. Hence, this approach cannot be applied directly to more general dubbing applications where source and target actors differ and data is relatively scarce.

2.2 Visual-Based Dubbing Techniques

Visual dubbing techniques can be classified into image-based or model-based approaches. Image-based techniques edit images directly in 2D image space. Geng et al. [2018] presented a warp-guided technique capable of controlling a single target image through a source driving video. The target image is warped according to the motion field of the driving video. Two generative adversarial networks are used, one for adding photorealistic fine visual details, and the other for synthesizing occluded regions such as the mouth interior. Wiles et al. [2018] presented X2Face, an approach for controlling a target video through a source video or audio. They propose two networks: the first projects the target video into an embedded face representation, and the second network estimates a driving vector that encodes the desired facial expressions, head pose and so on. While the approaches of Geng et al. and Wiles et al. produce

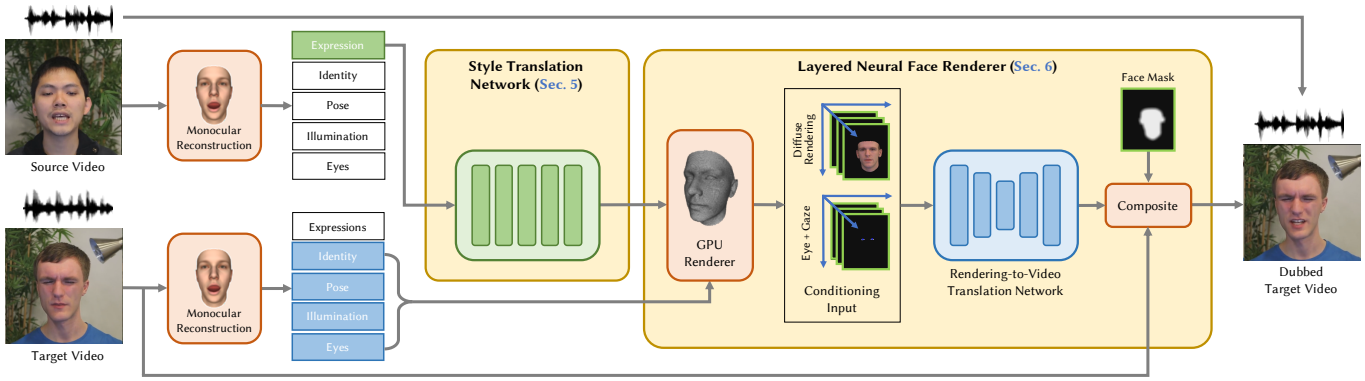

Fig. 2. Overview of our style-preserving visual dubbing approach. From left to right: First, we reconstruct the parameters of a 3D face model from the source and target input videos (Section 4). Next, we translate the source actor's facial expressions using our novel target-style preserving recurrent generative adversarial network (Section 5). Finally, we generate a photorealistic rendering of the dubbed target actor using a layer-based approach that composites a neural face rendering on top of dynamic video backgrounds (Section 6).

compelling results, the outputs often suffer from unnatural movements. In addition, neither approach is designed to maintain the target actor's style.

Model-based approaches rely on a parametric face model, namely a 3D deformable model. The Blanz and Vetter face model [2003; 1999] is commonly used to represent the identity geometry and albedo of the face. Facial expressions, including mouth movements, are usually modeled through blendshape parameters [Pighin et al. 1998]. For dubbing, the facial expressions of the source video are directly copied to the target video. Garrido et al. [2015] transfer the blendshape weights of the mouth region from the source to the target video by overlaying a rendered target face model. Finally, salient mouth movements, such as lip closure and opening, are imposed through the help of the dubbing audio track. For this, the bilabial consonants (/b/, /m/ and /p/) are detected from the audio track. Thies et al.'s Face2Face [2016] allows dubbing in real time from a monocular source video by overlaying a modified rendered face model on the target. Static skin texture is used and a data-driven approach synthesizes the mouth interior. Ma and Deng [2019] present an unpaired learning framework with cycle consistency for facial expression transfer. Unlike our approach, it transfers the same mouth expression from the source actor without considering the target's style. To this end, it introduces an additional lip correction term that simply measures the 3D distance of lip vertices.

Kim et al. [2018] presented Deep Video Portraits, a technique capable of producing high-quality photorealistic dubbing results. At first, a synthetic rendering of the target actor is produced, which captures the facial expressions of the source actor while maintaining the target actor's identity and pose. A conditional generative adversarial network translates the synthetic rendering into a photorealistic video frame. This approach is trained per target video. Nagano et al. [2018] proposed a similar approach, which however does not require identity-specific training. As a result, their approach can drive any still image by a given source video, but it only synthesizes the face region and not the hair. The still image is assumed to have a frontal perspective and a neutral pose. Even though model-based techniques provide full control over the target video, many suffer from audio-visual misalignments and are not designed to create high-quality videos with general backgrounds. In addition, model-based techniques often exhibit noticeable artifacts in synthesized mouth interiors [Garrido et al. 2015; Nagano et al. 2018; Thies et al. 2016] or dynamic skin textures [Thies et al. 2016]. Moreover, none of them maintain the style of the target actor.

2.3 Image-to-Image Translation

Learning-based image-to-image translation techniques have shown impressive results for a number of applications [e.g. Isola et al. 2017]. The core component is a conditional generative adversarial network that learns a mapping from the source to the target domain. This, however, requires paired training data, an assumption that is not easily satisfied. The introduction of cycle-consistency losses enabled learning from unpaired training data [Kim et al. 2017; Yi et al. 2017; Zhu et al. 2017] without explicit training pairs of source and target images. The cycle-consistency loss is defined such that a mapping from the source to the target followed by the inverse of this mapping should lead to the original source. This constraint is also applied in the opposite direction separately, i.e., from the target to the source. In our work, we utilize a cycle-consistency loss in model parameter space to train our visual dubbing framework on unsynchronized training data.

3 OVERVIEW

Different actors speak in different ways, using their own facial expressions. These person-specific idiosyncrasies need to be preserved during the visual dubbing process, hence one cannot simply copy the facial expressions from a source actor to a target actor as done in previous work. In order to achieve this, we propose to learn a *style-preserving* mapping between facial expressions in an unsupervised manner.

Our approach consists of three stages (Figure 2): monocular face reconstruction, style-preserving expression translation, and layered neural face rendering. The first stage of our approach registers a 3D face model to the source and target input videos (Section 4). This step reconstructs the facial expression parameters for every

video frame. The second stage of our approach is a novel style-preserving translation network (Section 5). We introduce a recurrent generative adversarial network that learns to transfer the source actor's expressions while maintaining the idiosyncrasies of a specific target actor. We train this network in an unsupervised manner on unpaired videos using cycle-consistency and mouth expression losses. The third stage of our approach is a new layered neural face renderer (Section 6) that generates photorealistic video frames from the style-translated expression parameters. We adopt the recent neural face rendering approach of Kim et al. [2018], and extend it for dynamic video backgrounds. Specifically, we introduce a soft face mask to blend the rendered photorealistic faces with the existing target video background layer.

4 3D FACE MODELING

We map the expression transfer problem from screen space to parameter space by registering a parametric 3D face model to every video frame. This later enables us to robustly learn a style-preserving expression mapping from the source to the target actor domains in an unsupervised manner.

We employ a parametric face model that encodes the head pose, face identity (geometry and appearance), and facial expression based on a low-dimensional vector. In more detail, we recover, for each frame f, the pose of the head $\mathbf{T} \in \mathrm{SE}(3)$, face geometry $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, face reflectance $\boldsymbol{\beta} \in \mathbb{R}^{80}$, face expression $\boldsymbol{\delta} \in \mathbb{R}^{64}$, and spherical-harmonics illumination $\boldsymbol{\gamma} \in \mathbb{R}^{27}$. For dimensionality reduction, the geometry and appearance bases have been computed based on 200 high-quality scans [Blanz and Vetter 1999] using principal component analysis (PCA). The low-dimensional expression subspace has been computed via a PCA of the facial blendshapes of Cao et al. [2014] and Alexander et al. [2010]. We recover all parameters from monocular video based on an optimization-based 3D face reconstruction and tracking approach inspired by Garrido et al. [2016] and Thies et al. [2016]. The energy is composed of a dense color alignment term between the input image and the rendered model, a sparse alignment term based on automatically detected facial landmarks [Saragih et al. 2011], and a statistical regularizer. The facial landmark tracker also recovers the 2D image position of the pupils of the left eye, $\mathbf{e}_\mathrm{l} \in \mathbb{R}^2$, and right eye, $\mathbf{e}_\mathrm{r} \in \mathbb{R}^2$. This procedure lets us fully automatically annotate each video frame f with a low-dimensional parameter vector $\mathbf{p}_f \in \mathbb{R}^{261}$. Of specific importance for us in the later processing steps are the recovered expression parameters $\boldsymbol{\delta}$. In the following, we show how to robustly learn a style-preserving mapping between the expression parameters of two different actors without requiring paired training data.

5 STYLE TRANSLATION NETWORK

We propose a style translation network that learns a mapping from the distribution of the source actor expressions to the distribution of the target actor expressions, and vice versa, using cycle consistency. Our approach is inspired by recent techniques for unpaired image-to-image translation [e.g. Kim et al. 2017; Yi et al. 2017; Zhu et al. 2017], and shares a similar high-level design: We employ a generative adversarial network with two generator networks and two discriminator networks, which are trained in an unsupervised fashion from unpaired training data using cycle-consistency, adversarial and mouth expression losses. The generator networks translate from the source to the target domain, and vice versa, while there is a discriminator network for each of the two domains. However, everything else is different as we learn the translation of temporal facial expression parameters, i.e., multiple vectors $\boldsymbol{\delta}$ corresponding to multiple video frames. Specifically, we propose a *recurrent* generative adversarial network to encode the temporal dynamics of the learned distribution of facial expressions using long short-term memory (LSTM) units [Hochreiter and Schmidhuber 1997].

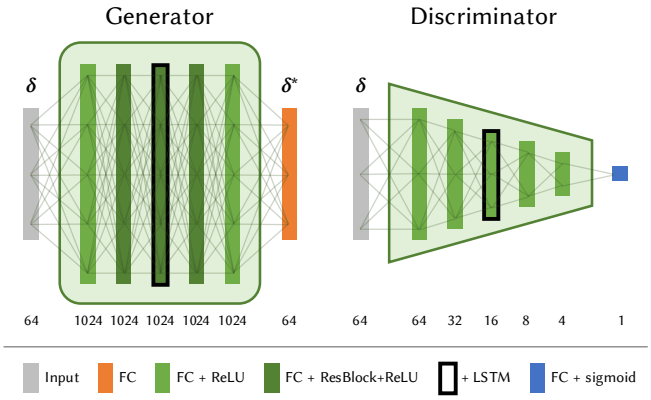

Fig. 3. Illustration of the architectures of the generator and discriminator networks as part of our style translation network (Section 5).

5.1 Network Architecture

The input to both the generator and the discriminator networks is a tuple $(\boldsymbol{\delta}_{t-N+1}, \ldots, \boldsymbol{\delta}_t)$ of $N = 7$ facial expression parameters $\boldsymbol{\delta}_f \in \mathbb{R}^{64}$. The tuple comprises the six frames before frame t and the current frame t. Each parameter vector $\boldsymbol{\delta}$ is calculated using the approach in Section 4, and we normalize each expression component to zero mean and unit variance (per video). We illustrate the architecture of our networks in Figure 3.

Generators. The generator consists of five fully-connected layers, each containing 1024 nodes with ReLU activations. We use residual blocks [He et al. 2016] in the middle three layers to facilitate the learning of deviations from the identity expression translation function. This effectively captures the difference between the target and source distributions. We further implement LSTM units in the middle layer of the generator to encode the temporal dynamics of facial expressions. This is defined over $N = 7$ consecutive frames, specifically the current frame and the six preceding frames. A final fully-connected layer of 64 nodes outputs the parameters of the translated expression coefficients, without any activation function.

Discriminators. The discriminator network comprises five fully-connected layers with ReLU activations. The first layer has 64 nodes, matching the size of the input, and each subsequent layer has half as many nodes. The middle layer also implements LSTM units to consider the temporal dynamics of facial expressions. The output node is fully connected to the previous layer and uses a sigmoid

activation to produce a value in the unit range. Given a tuple of N facial expression vectors as input, the discriminator produces an output with N real values in the unit range.

5.2 Training Loss

To train our style translation network in an unsupervised manner, we combine three losses into our objective function:

$$L = \lambda_{\text{cc}} L_{\text{cc}} + \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{me}} L_{\text{me}}. \tag{1}$$

Here, L_{cc} is the cycle-consistency loss that enables training with unpaired training data, L_{adv} is the adversarial loss that encourages the output of the generator to better match the target domain, and L_{me} is a novel cosine mouth expression loss that promotes corresponding mouth expressions, such as mouth closure, between source and target. Each loss is weighted by a corresponding coefficient $\lambda_\bullet$.

5.2.1 Cycle-Consistency Loss. There are two generators in our network that translate facial expression parameters from the distribution of source actor expressions, $\mathcal{S}$, to the distribution of target actor expressions, $\mathcal{T}$, and vice versa. $\text{G}_{\mathcal{S}\to\mathcal{T}}$ denotes the translation of the source actor expression $\mathbf{s} = (\boldsymbol{\delta}_{t-N+1}, \ldots, \boldsymbol{\delta}_t) \in \mathcal{S}$ to a target actor expression $\text{G}_{\mathcal{S}\to\mathcal{T}}(\mathbf{s}) \in \mathcal{T}$. This captures the spatial and temporal co-activations of the facial expressions while preserving the style of the target. On the other hand, $\text{G}_{\mathcal{T}\to\mathcal{S}}$ is the mapping in the opposite direction, from the target distribution $\mathbf{t} \in \mathcal{T}$ into the source distribution $\text{G}_{\mathcal{T}\to\mathcal{S}}(\mathbf{t}) \in \mathcal{S}$. Composing both generators, and measuring the distance to the starting point, results in the cycle-consistency loss

$$L_{\text{cc}} = \left\|\text{G}_{\mathcal{T}\to\mathcal{S}}(\text{G}_{\mathcal{S}\to\mathcal{T}}(\mathbf{s})) - \mathbf{s}\right\|_1 + \left\|\text{G}_{\mathcal{S}\to\mathcal{T}}(\text{G}_{\mathcal{T}\to\mathcal{S}}(\mathbf{t})) - \mathbf{t}\right\|_1, \tag{2}$$

where $\mathbf{s} \in \mathcal{S}$ and $\mathbf{t} \in \mathcal{T}$ are unpaired training samples, and we use an ℓ_1-loss to measure similarity. Using both generators with cycle consistency allows us to train our approach in an unsupervised manner, with unpaired data. This is important since it is challenging to obtain paired data (time-synchronized face video across different languages) for our problem of visual dubbing from one language to another.

5.2.2 Adversarial Loss. Both generators are accompanied by discriminators, $\text{D}_{\mathcal{S}}$ and $\text{D}_{\mathcal{T}}$, which correspond to the source and target domains, respectively. The discriminators work towards getting better in classifying the generated result as either real or synthetic, while the generators aim to fool the discriminators by improving the quality of their output. The input to each discriminator is a temporal vector of N facial expression vectors, and its output is a vector of N real numbers, corresponding to the individual input vectors. This, in combination with the LSTM units, allows us to better capture the temporal correlations between the examined distributions. We define the adversarial loss in a bidirectional manner as follows:

$$\begin{aligned} L_{\text{adv}} = \log \frac{\|\text{D}_{\mathcal{T}}(\mathbf{t})\|_1}{N} + \log\left(1 - \frac{\|\text{D}_{\mathcal{T}}(\text{G}_{\mathcal{S}\to\mathcal{T}}(\mathbf{s})\|_1}{N}\right) \\ + \log \frac{\|\text{D}_{\mathcal{S}}(\mathbf{s})\|_1}{N} + \log\left(1 - \frac{\|\text{D}_{\mathcal{S}}(\text{G}_{\mathcal{T}\to\mathcal{S}}(\mathbf{t})\|_1}{N}\right). \end{aligned} \tag{3}$$

As before, $\mathbf{s} \in \mathcal{S}$ and $\mathbf{t} \in \mathcal{T}$ are unpaired training samples.

5.2.3 Cosine Mouth Expression Loss. The unpaired training using cycle-consistency and adversarial losses does not always preserve important mouth expressions, such as opening or closing. We therefore introduce an additional loss to encourage the correct translation of these important mouth expressions. Specifically, we use a cosine loss on the ten mouth-specific facial expressions between the source and target domains. The cosine loss is more effective in aligning the mouth-related source and target expressions with different magnitudes, corresponding to different styles, than the Euclidean loss. This encourages our network to maintain correspondence in mouth expressions without unduly constraining the strength of these expressions, as this would for example alter the intensity of smiles between actors. We again use a symmetric loss,

$$L_{\text{me}} = L_{\cos}(\mathbf{s}, \text{G}_{\mathcal{S}\to\mathcal{T}}(\mathbf{s})) + L_{\cos}(\mathbf{t}, \text{G}_{\mathcal{T}\to\mathcal{S}}(\mathbf{t})), \tag{4}$$

where $L_{\cos}$ computes the mean cosine distance between the mouth-specific expressions over all time steps:

$$L_{\cos}(\mathbf{s}, \mathbf{t}) = \frac{1}{N} \sum_{n=1}^{N} \frac{\boldsymbol{\mu}(\mathbf{s}_n) \cdot \boldsymbol{\mu}(\mathbf{t}_n)}{\|\boldsymbol{\mu}(\mathbf{s}_n)\|_2 \cdot \|\boldsymbol{\mu}(\mathbf{t}_n)\|_2}. \tag{5}$$

Here, we use the notation $\mathbf{s}_n$ to select the n^{th} element of the tuple $\mathbf{s}$, which is the facial expression vector $\boldsymbol{\delta}$ of a source actor, and from which the function $\boldsymbol{\mu}(\cdot)$ selects the ten mouth-specific expression coefficients. We select the ten coefficients with the largest mouth expression variation by visual inspection of the rendered PCA basis.

5.3 Network Training

Our training dataset is a collection of sequential expression parameters from individual videos, recovered using the monocular 3D face reconstruction approach in Section 4. We found that facial expression styles consistently captured by approximately five-minute-long videos are typically sufficient to train our style translation network. As preprocessing, we normalize each expression coefficient to zero mean and unit variance, and then extract sliding windows of size $N = 7$ frames. We balance the influence of loss functions in Equation 1 using $\lambda_{\text{cc}} = 10$, $\lambda_{\text{adv}} = 1$ and $\lambda_{\text{me}} = 5$. The loss is minimized using the Adam solver [Kingma and Ba 2015] with an initial learning rate of 0.0001 and an exponential decay rate of 0.5. During back-propagation for the LSTM units, we apply a gradient norm clipping operation to avoid exploding gradients [Pascanu et al. 2013]. We implement our network using the TensorFlow deep learning library [Abadi et al. 2015]; training typically converges within 25 epochs.

6 NEURAL FACE RENDERER

The final step of our approach is to synthesize a photorealistic portrait video from a sequence of face model parameters (see Section 4) that correspond to the dubbed target actor. Our approach builds on recent advances in neural rendering for creating high-fidelity visual dubbing results. Specifically, we extend deep video portraits [Kim et al. 2018], which assumes a static video background, to support the dynamic video backgrounds found in feature films, television series, and many real-world videos. Considering our target application of dubbing, the only area that needs to be modified is the face interior of the target actor, as we would like to preserve the remainder of the target actor's performance. Unlike Kim et al., we therefore restrict

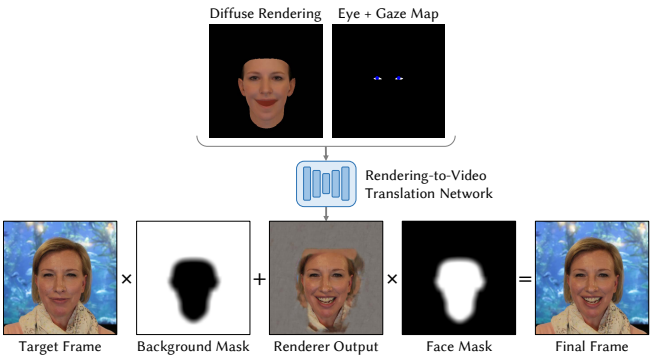

Fig. 4. We use a layer-based neural face renderer that composes the neural face rendering onto the dynamic video background using a soft face mask.

the neural renderer to the face region and composite the predicted face rendering over the target video. This leaves any potentially dynamic background intact in the final face rendering.

Figure 4 shows the work flow of our layer-based renderer. We first rasterize diffusely shaded renderings of the face model, as well as eye maps with proxy pupils using the standard graphics pipeline. Unlike Kim et al. [2018], we do not render texture coordinates as we found them to be not necessary. We feed these images into Kim et al.'s rendering-to-video translation network [2018], which we use at 512×512 pixel resolution by default. To focus the network on the face region, we apply the face mask to the predicted image and the ground truth before passing them to the discriminator and computing the adversarial loss. The employed additional per-pixel ℓ_1-loss with respect to ground truth is only computed within the face mask region, and the relative weight between the ℓ_1 and adversarial losses is set to a ratio of 100:1, as in Kim et al. [2018]. The face mask covers the face interior between the ears, from the forehead down to the laryngeal prominence (Adam's apple). Finally, we composite the predicted face over the current target video frame. To achieve seamless blending, we erode the binary face mask to reduce its size slightly, and then smooth its boundaries with a Gaussian filter. The face rendering is then composited onto the original target video using the soft face mask. Since we did not modify the head pose, the composition appears seamless in most cases. For better temporal consistency in the generated results, we process a video using a moving window. The input to our network is a space-time tensor defined over 7 frames (including the current frame as the last frame). The tensor therefore has a dimension of $W \times H \times (7 \cdot 2 \cdot 3)$, i.e., stacking all 7 conditioning inputs (2 images with 3 color channels each).

7 RESULTS

We demonstrate our style-preserving video dubbing approach, perform a qualitative and quantitative evaluation (user study), and thoroughly compare to the state of the art in audio-based and video-based dubbing. Please see our supplemental video for audio-visual results and comparisons. We start by giving an overview of the used sequences, and discuss the runtime requirements of our approach.

Datasets. We tested our approach on a diverse set of 11 source and 12 target sequences, which are detailed in Table 4. The average length of both the source and target sequences is on average five minutes. In total, we dubbed over 50 minutes of video footage with our approach. The source and target sequences show different people, all with their own person-specific idiosyncrasies and style, in front of a large variety of backgrounds, both static as well as dynamic. We also show dubbing results between different languages, such as German-to-English dubbing. The resolution of all produced videos is 512×512 pixels.

Runtime Requirements. Dense 3D face reconstruction and tracking takes 250 ms per video frame. Training our style-preserving expression mapping takes 8 hours per sequence on an Nvidia Tesla V100. At test time, applying the mapping takes 952 ms per video frame and our neural renderer requires additional 224 ms to produce the final output.

Training and Testing. We learn the style-preserving mapping between the source and target sequences in an unsupervised manner. An input video is split into training and test sets. We usually use the first 7,500 frames for training. At test time, we feed the the rest of the frames to our system as illustrated in Figure 2. Our style translation network is source-to-target specific. Therefore, we retrain the network with different video pairs to handle different styles or identities. Our neural renderer is also person-specific and trained with all the frames from a target video without any split.

7.1 Visual Dubbing

Visual dubbing is an approach to change the mouth motion of a target actor, such that it matches the voice of a dubbing actor that speaks in a foreign language. One example of this is dubbing an English movie to German. Existing visual dubbing approaches directly copy the mouth motion of the dubbing actor to the target. While this leads to good audio-visual alignment, it also removes the person-specific idiosyncrasies and the style of the target actor, and makes the target actor's face move unnaturally like the source actor. Our style-preserving video dubbing approach enables to achieve good audio-visual alignment, while also preserving the idiosyncrasies and the style of the target actor, see Figure 5. For example, we dub a very expressive actor speaking in German to English based on a neutrally speaking dubbing actor. As can be seen, our approach is able to preserve the idiosyncrasies and the style of the target actor well. For the female target actor (Figure 5, left), our technique maintained her wide and happy mouth openings. For the the male target actor (Figure 5, right), we maintained his near-closed eyes while modifying his mouth movements to match the dubbing track. For the full result videos, we refer to the supplemental video.

Our style-preserving dubbing approach can also be used to handle cases when the source actor has an expressive style. This is a more challenging problem as the translation network needs to remove the strong source style and replace it with the different target style. Figure 6 shows the results from expressive styles. On the left, the source actor is smiling while the target is angry. On the right, the source actor is angry while the target is smiling. In both examples, our approach captures the source mouth movements and maintain the target style. These examples also demonstrate our visual dubbing

Fig. 5. We demonstrate our style-preserving visual dubbing approach for two German-speaking 'Target' actors with strong expressive styles (in the middle row), who we are dubbing into English using neutral 'Source' actors (in the top row). As can be seen, our results (in the bottom row) do preserve the idiosyncrasies and the style of the target actors well. Note that all videos feature dynamic backgrounds.

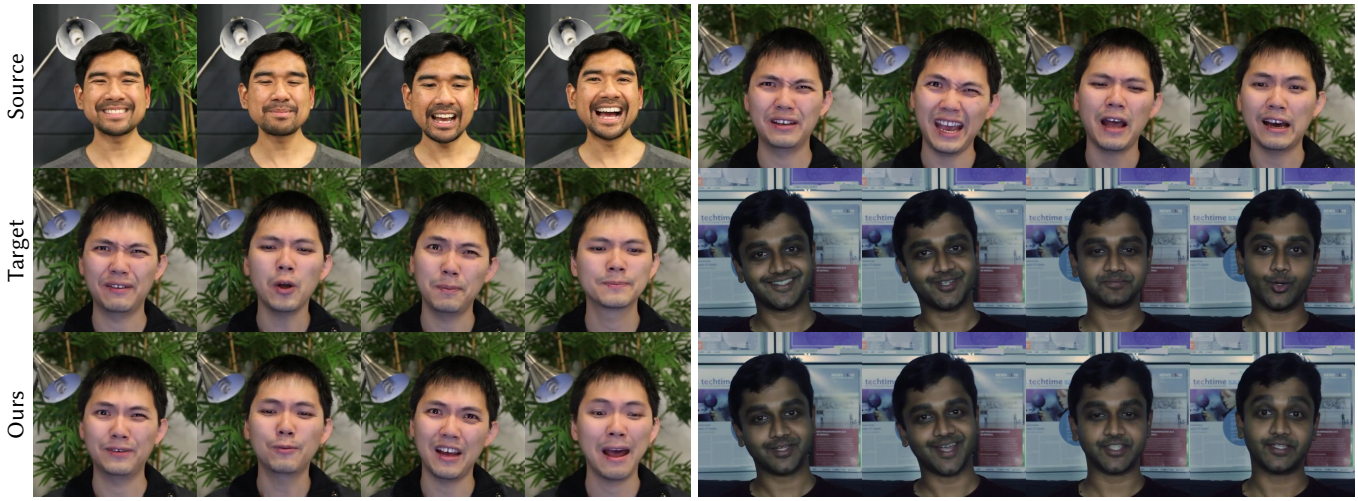

Fig. 6. Our dubbing technique can handle expressive source styles. As can be seen, our approach is able to transfer the source facial expressions, while preserving the idiosyncrasies and style of the target actor. We demonstrate English-to-Indonesian dubbing (left) and Kannada-to-English dubbing (right).

method in other languages: English to Indonesian on the left and Kannada to English on the right.

Our dubbing technique naturally handles similar expressions between the source and target actors, as we show in Figure 7. We note that similar expressions still appear differently due to person-specific styles. Our technique dubs the target actor while maintaining his/her specific style. We include additional results in our supplemental video.

7.2 Comparisons to the State of the Art

We perform extensive comparisons to the current state of the art in visual dubbing, facial reenactment and audio-based reenactment. More specifically, we compare to the VDub [Garrido et al. 2015], Deep Video Portraits [Kim et al. 2018], and Audio2Obama [Suwajanakorn et al. 2017] approaches. We also perform a baseline comparison with the unpaired image-to-image translation approaches CycleGAN [Zhu et al. 2017] and UNIT [Liu et al. 2017].

First, we compare to the state-of-the-art audio-visual VDub approach [Garrido et al. 2015] in Figure 9. The VDub approach leverages the audio channel to better align the visual content with lip closure events, but is not able to preserve the style of the target actor. In contrast, our style-preserving visual dubbing approach enables us to maintain the style of the target actor and achieve a good audio-visual alignment. In addition, our learning-based approach synthesizes a higher quality mouth interior than the model-based VDub approach that only renders a coarse textured teeth proxy.

Fig. 7. Our dubbing technique naturally supports source and target actors with similar expressions. In these examples, both actors are either smiling (left) or neutral (right). In each case, our dubbing result (bottom row) maintains the specific style of each target actor. On the left, the target has an expressive smile with a wide mouth opening, which is maintained in our result. On the right, the target actor speaks in a narrower mouth shape than the source. This target style is preserved in our result. For these and additional video results, please refer to the supplemental video.

Fig. 8. Comparison to a variant of Deep Video Portraits [Kim et al. 2018], i.e. our approach without the style translation network. Our approach enables us to preserve the style and idiosyncrasies of each target actor. This is in contrast to other dubbing approaches, such as Kim et al. [2018], that copy expressions directly and are hence not style-preserving. On the left, our results maintain the wide-open mouth of the target, while the variant of Kim et al. narrows it to match the source style. On the right, our technique captures the target style more naturally, while the variant of Kim et al. generates artifacts (see arrows).

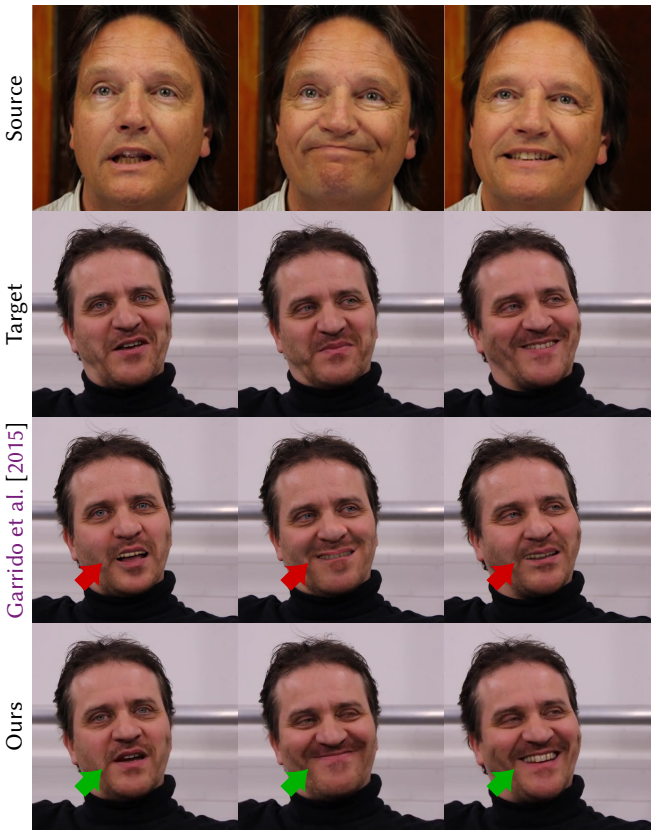

Fig. 9. Comparison to Garrido et al. [2015]. Our approach is able to preserve the style and idiosyncrasies of the target actor, while these get lost with the approach of Garrido et al. [2015]. In addition, our approach synthesizes a higher quality and more realistic mouth interior.

Fig. 10. Comparison to the audio-based dubbing approach of Suwajanakorn et al. [2017]. Their results are not always consistent with the source actor, with the mouth being open or closed when it should not be (see red arrows). In addition, this approach does not explicitly deal with style, as it can only perform self-reenactment. Videos: The White House (public domain).

In Figure 8, we compare our approach to a variant of Deep Video Portraits [Kim et al. 2018], a learning-based facial reenactment approach that also supports visual dubbing. Specifically, we use our approach without the style translation network, i.e., we pass the source facial expressions directly to our layer-based renderer, which handles dynamic backgrounds unlike Deep Video Portraits. Unlike other dubbing approaches, our style-preserving visual dubbing approach enables us to maintain the style and idiosyncrasies of the target actor. For more detail, please refer to the supplemental video.

We compare to the audio-based Audio2Obama facial reenactment approach [Suwajanakorn et al. 2017] in Figure 10. Their approach can control a virtual version of Barack Obama based on a new audio clip of himself. Note that since this approach only works from-Obama-to-Obama, it does not explicitly deal with style, since there is no style to preserve in this setting. Nevertheless, our visual dubbing approach leads to a more faithful reproduction of the actual visual content in the dubbing sequence, while the audio-based approach misses mouth motions that are uncorrelated with the audio track. In addition, Audio2Obama is tailored for Barack Obama and trained on 17 hours of his speeches. Our technique, however, is trained on only 2 minutes from just a single video. For video results, please see our supplemental video.

Finally, we perform a baseline comparison to CycleGAN [Zhu et al. 2017] and UNIT [Liu et al. 2017], which are unsupervised image-to-image translation techniques. These image-based techniques are not able to fully disentangle styles from lip motion and dynamic backgrounds. Unsupervised translation is a much harder problem in the image domain, due to the higher dimensionality of the problem. This problem is solved by our approach that transfers the unsupervised learning problem into parameter space. In this space, the problem is lower dimensional and the mapping can be found more robustly, as illustrated in Figure 11.

7.3 Ablation Study

We also performed an ablation study to evaluate the components of our novel style-preserving visual dubbing approach. A quantitative ablation study is challenging to perform, since there is no ground-truth metric available for style-preserving dubbing. Therefore, we perform a qualitative ablation study in Figure 12 and in the supplemental video. We evaluate the influence of the LSTM unit, the cycle-consistency loss, the cosine mouth expression loss, and the face attention map used during training. Figure 12 shows that the LSTM is important in producing temporally coherent results with

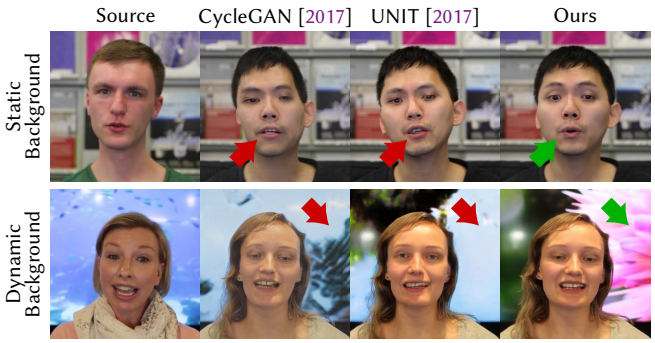

Fig. 11. Comparison to the unpaired image-to-image translation techniques CycleGAN [Zhu et al. 2017] and UNIT [Liu et al. 2017]. Unsupervised transfer is a much harder problem in the image domain, due to the higher dimensionality of the problem. We formulate the unsupervised transfer problem in parameter space, which leads to higher quality results.

better lip syncing. removing the cycle-consistency loss has the impact of distorting the face in a highly unnatural manner. The cosine mouth expression loss better captures the audio-visual alignment of the mouth region. Removing the background attention map leads to noticeable artifacts in the final generated background. This is best seen in the supplementary video.

7.4 User Study

We performed an extensive web-based user study to quantitatively evaluate the quality of the generated visual dubbing results. We conducted two experiments using a collection of 12 short video clips (3–8 seconds): a subjective rating task and a pairwise comparison task using two-alternative forced choice (2AFC). We prepared all results at 512×512 pixels resolution and showed either one or two videos side by side, depending on the experiment, but always with the target audio track. Most video clips are from videos we recorded ourselves (e.g. see Figures 5 to 9); they show different people talking in multiple languages with a range of facial expressions, including sarcastic, smiling and squinting. We ensured that these clips have an unbiased distribution of source and target expressions, so as to not influence users adversely, e.g. with more favorable target expressions. For these video clips, we compare to the 'naïve dubbing' baseline of combining the source audio track with the target video track without modifying the video, as well as our approach without the style translation network. We also used a professionally dubbed video created by Garrido et al. [2015], i.e., a studio actor's text and speed was optimally aligned to the existing video. We compare to both the professional dubbing and Garrido et al.'s approach. We recruited 50 anonymous participants who took on average 13.7 minutes to complete our study.

The results in Table 1 show that our approach (60/53% 'natural' rating) clearly outperforms naïve dubbing (8%) as well as our approach without style translation (54/43%), and it achieved the quality of professional dubbing (54%), all while being fully automatic. The pairwise comparison results in Table 2 provide three insights: 1. There is a clear preference (>80%) of our style-preserving dubbing approach on video clips 4 and 5, in which a neutral source actor

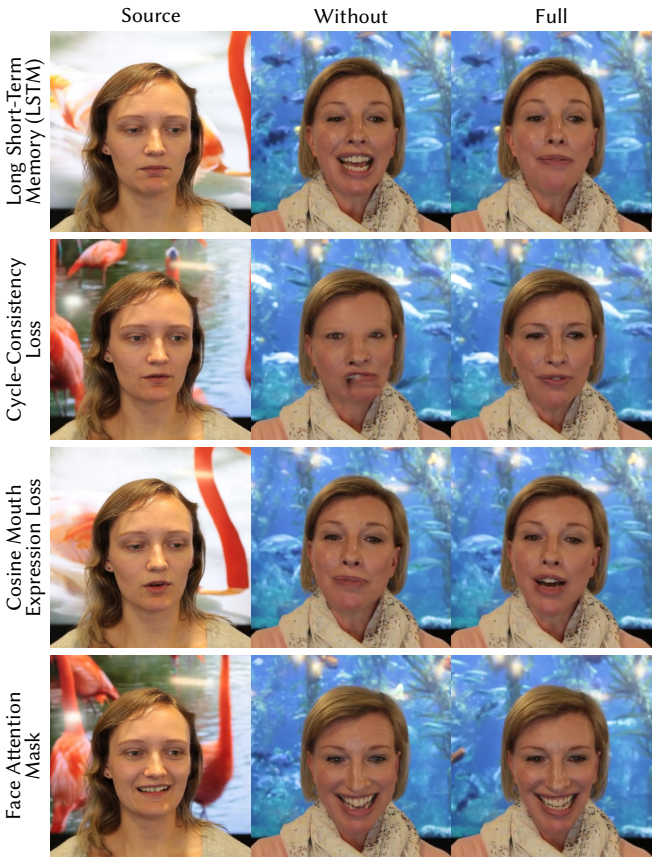

Fig. 12. We perform an ablation study by disabling each component of our method to evaluate their impact on the end results. The LSTM unit better captures the temporal changes of facial expressions, leading to better lip-syncing and more plausible style translation. Without cycle consistency, the style mapping function becomes ill-posed, and thus is prone to generate invalid expressions. This generates significant artifacts in the final rendering. The cosine mouth expression loss leads to accurate lip synchronization, keeping the strength of mouth expressions close to the target actor style. The face mask in our neural renderer improves the visual quality of synthesized outputs, enabling dynamic backgrounds and artifact-free teeth reconstruction. The improvements are best visible in our video.

dubs a smiling target actor (shown in Figure 1), so the lack of style preservation is immediately obvious. 2. Preferences for the other video clips are mixed (49% mean), perhaps as the target actor style was not shown for comparison. 3. While our approach is clearly preferred over naïve dubbing (89% mean preference), there are 2 minor outliers (<80%) where lip motions somewhat aligned to the audio by chance. Finally, we performed pairwise comparisons on Garrido et al.'s professionally dubbed video in Table 3, which shows that users found our style-preserving dubbing result more natural than VDub [2015] (73%) and our approach without style preservation (78%). 41% of users preferred our result to the professional dubbing result, compared to only 26% for without style translation and 21% for Garrido et al. [2015]. This is a clear improvement over the state of the art.

Table 1. User study results ($n = 50$) in response to the statement "This video clip looks natural to me", from "−−" (*strongly disagree*) to "++" (*strongly agree*). **Top:** Mean of 9 video clips with a variety of source and target actors. **Bottom:** Mean of 3 video clips from a professionally dubbed video.

	−−	−	o	+	++	agree
Naïve dubbing	54	33	5	8	0	8%
Ours without style translation	9	25	12	36	18	54%
Our with style translation	5	23	11	44	16	60%
Garrido et al. [2015]	34	27	8	25	5	31%
Ours without style translation	10	32	15	33	9	43%
Our with style translation	3	33	11	44	9	53%
Professional dubbing	9	27	10	44	10	54%

Table 2. User study results for two-alternative forced choice (2AFC) on 9 video clips ($n = 50$). Row "A > B" shows %users who found A more natural.

	1	2	3	4	5	6	7	8	9	mean
Ours > w/o style translation	52	50	50	84	82	50	38	52	52	57%
Ours > naïve dubbing	96	84	92	94	92	98	98	74	76	89%

Table 3. User study results for two-alternative forced choice (2AFC) for a professionally dubbed video ($n = 50$, mean of 3 clips).

% row preferred over column		VDub	w/o	ours	prof
Garrido et al. [2015]	VDub	—	39	27	21
Ours without style translation	w/o	61	—	22	26
Our with style translation	ours	73	78	—	41
Professional dubbing	prof	79	74	59	—

To better understand *why* participants found our results with style translation more natural, we considered the impact of potential discrepancies between source and target actor performances, expressions and styles. The preference for our results is not simply due to a more favorable target style: both source and target videos of the professionally dubbed video (Table 3) are neutral, yet users still prefer style translation 78% of the time. The remaining 9 video clips are 2× smile-to-smile, 3× neutral-to-smile, 2× neutral-to-sarcastic, 2× neutral-to-squint, which is unbiased overall. The target training video is also naturally more representative of the target actor's expressions and style, and might lack specific source expressions, such as a broad smile. This favors our style translation approach. In dubbing applications, the source and target videos are roughly aligned in time, which avoids visually incoherent source expression and target pose in most cases. In practice, we observed that our approach can still cope with small misalignments and produces good results.

7.5 Style Interpolation

In addition to style-preserving visual dubbing, our approach can also be used to change the style of the target actor in interesting and meaningful ways. For example, we can smoothly blend between the style of the source and the target actor, see Figure 13, while controlling the target sequence. This can, for example, be used to make the target a bit more happy, neutral or sad, by mixing in some

Fig. 13. Seamless interpolation between the source and target actor style. We linearly blend the original and translated source expressions using the weight $\alpha \in [0, 1]$, and visualize the result with our neural renderer. Note that the style interpolation is achieved with synchronized lip motions.

Fig. 14. Our method might generate implausible facial expressions such as one-eyed twitching when the distributions of the facial expressions of the source and target domains are far apart. This could be avoided by incorporating additional constraints concerning facial anatomy. The temporal artifact is visible in the supplemental video.

of the style of the actor in the source sequence. One could imagine using this technique as a postprocessing step in a movie production to slightly adjust the tone of an already recorded performance.

8 DISCUSSION

In this work, we have demonstrated high-quality style-preserving visual dubbing results for a large variety of sequences. Our approach makes a step towards further simplifying localization of media, in particular movies and TV content, to different countries and languages. Nevertheless, our approach has a few limitations that can be tackled in the future.

Monocular face reconstruction is an extremely challenging problem and may fail for extreme illumination conditions or head poses, such as are often observed in two person dialogue shots. In these cases, the facial expressions can not be robustly recovered and thus the neural face renderer can not be reliably trained. More robust face reconstruction techniques could alleviate this problem in the future. Similar to many other data-driven techniques, training our neural face renderer requires a sufficiently large training corpus. Generalizing across subjects, to enable dubbing in settings where only a short video clip of the target actor is available, is an open challenge. Related to this, our approach works well inside the span of the training corpus, but generalization to unseen expressions is hard, e.g., synthesizing an extreme opening of the mouth, if this has not been observed before, might lead to visual artifacts.

The unsupervised training of the style translation might generate implausible facial expressions if the source and target distributions are too different. We noticed this is mostly the case when the source dubber has an extreme style, such as squinting or twitching. In Figure 14, the extreme twitching of the source actor produced visual

artifacts in the dubbed result. These artifacts could potentially be reduced by incorporating additional constraints on facial anatomy during training. Out of 19 videos in our dataset (Table 4), only 2 contain such extreme styles (11%). Differences in ethnicity or gender between the source and target may further contribute to differences in styles, which we leave for future work. We acknowledge minor visual artifacts in the final renderings but consider improvements in the rendering quality orthogonal to our main contribution of style-preserving dubbing.

9 CONCLUSION

We have presented the first style-preserving visual dubbing approach that maintains the signature style of the target actor including their person-specific idiosyncrasies. At the core of our approach is an unpaired temporal parameter-to-parameter translation network that can be trained in an unsupervised manner using cycle-consistency and mouth expression losses. Afterwards, photorealistic video frames are synthesized using a layered neural face renderer. Our results demonstrate that a large variety of source and target expressions, across subjects from different ethnicities and speaking different languages, can be handled well. We see our approach as a step towards solving the important problem of video dubbing and we hope it will inspire more research in this direction.

ACKNOWLEDGMENTS

We are grateful to all our actors and the reviewers for their valuable feedback. We thank True-VisionSolutions Pty Ltd for kindly providing the 2D face tracker and Adobe for a Premiere Pro CC license. This work was supported by ERC Consolidator Grant 4DRepLy (770784), the Max Planck Center for Visual Computing and Communications (MPC-VCC), RCUK grant CAMERA (EP/M023281/1), and an EPSRC-UKRI Innovation Fellowship (EP/S001050/1).

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. 2010. The Digital Emily Project: Achieving a Photorealistic Digital Actor. *IEEE Computer Graphics and Applications* 30, 4 (July/August 2010), 20–31. https://doi.org/10.1109/MCG.2010.65

Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. 2003. Reanimating Faces in Images and Video. *Computer Graphics Forum* 22, 3 (September 2003), 641–650. https://doi.org/10.1111/1467-8659.t01-1-00712

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *SIGGRAPH*. 187–194. https://doi.org/10.1145/311535.311556

Matthew Brand. 1999. Voice Puppetry. In *SIGGRAPH*. 21–28. https://doi.org/10.1145/311535.311537

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (March 2014), 413–425. https://doi.org/10.1109/TVCG.2013.249

Young-Woon Cha, True Price, Zhen Wei, Xinran Lu, Nicholas Rewkowski, Rohan Chabra, Zihe Qin, Hyounghun Kim, Zhaoqi Su, Yebin Liu, Adrian Ilie, Andrei State, Zhenlin Xu, Jan-Michael Frahm, and Henry Fuchs. 2018. Towards Fully Mobile 3D Face, Body, and Environment Capture Using Only Head-worn Cameras. *IEEE Transactions on Visualization and Computer Graphics* 24, 11 (November 2018), 2993–3004. https://doi.org/10.1109/TVCG.2018.2868527

Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that?. In *British Machine Vision Conference (BMVC)*.

Zhigang Deng and Ulrich Neumann. 2006. eFASE: Expressive Facial Animation Synthesis and Editing with Phoneme-isomap Controls. In *Symposium on Computer Animation (SCA)*. 251–260.

Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2015. VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track. *Computer Graphics Forum* 34, 2 (May 2015), 193–204. https://doi.org/10.1111/cgf.12552

Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *ACM Transactions on Graphics* 35, 3 (June 2016), 28:1–15. https://doi.org/10.1145/2890493

Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-guided GANs for Single-photo Facial Animation. *ACM Transactions on Graphics* 37, 6 (November 2018), 231:1–12. https://doi.org/10.1145/3272127.3275043

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (November 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven Facial Animation by Joint End-to-end Learning of Pose and Emotion. *ACM Transactions on Graphics* 36, 4 (July 2017), 94:1–12. https://doi.org/10.1145/3072959.3073658

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Transactions on Graphics* 37, 4 (August 2018), 163:1–14. https://doi.org/10.1145/3197517.3201283

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/1703.05192

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

Sumedha Kshirsagar and Nadia Magnenat-Thalmann. 2003. Visyllable Based Speech Animation. *Computer Graphics Forum* 22, 3 (September 2003), 631–639. https://doi.org/10.1111/1467-8659.t01-2-00711

Bertrand Le Goff, Thierry Guiard-Marigny, Michael M. Cohen, and Christian Benoît. 1994. Real-time analysis-synthesis and intelligibility of talking faces. In *SSW*. https://www.isca-speech.org/archive_open/ssw2/ssw2_053.html

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. In *Advances in Neural Information Processing Systems (NIPS)*. https://github.com/mingyuliutw/unit

Jiyong Ma, Ron Cole, Bryan Pellom, Wayne Ward, and Barbara Wise. 2006. Accurate visible speech synthesis based on concatenating variable length motion capture data. *IEEE Transactions on Visualization and Computer Graphics* 12, 2 (March 2006), 266–276. https://doi.org/10.1109/TVCG.2006.18

Luming Ma and Zhigang Deng. 2019. Real-Time Facial Expression Transformation for Monocular RGB Video. *Computer Graphics Forum* 38, 1 (2019), 470–481. https://doi.org/10.1111/cgf.13586

Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 2018. paGAN: Real-time Avatars Using Dynamic Textures. *ACM Transactions on Graphics* 37, 6 (November 2018), 258:1–12. https://doi.org/10.1145/3272127.3275075

Elmer Owens and Barbara Blazek. 1985. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech, Language, and Hearing Research* 28, 3 (September 1985), 381–393. https://doi.org/10.1044/jshr.2803.381

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training Recurrent Neural Networks. In *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/1211.5063

Hai X. Pham, Samuel Cheung, and Vladimir Pavlovic. 2017. Speech-Driven 3D Facial Animation With Implicit Emotional Awareness: A Deep Learning Approach. In *CVPR Workshops*. https://doi.org/10.1109/CVPRW.2017.287

Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. 1998. Synthesizing Realistic Facial Expressions from Photographs. In *SIGGRAPH*. 75–84. https://doi.org/10.1145/280814.280825

Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2011. Deformable Model Fitting by Regularized Landmark Mean-Shift. *International Journal of Computer Vision* 91, 2

(2011), 200–215. https://doi.org/10.1007/s11263-010-0380-4

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Transactions on Graphics* 36, 4 (July 2017), 95:1–13. https://doi.org/10.1145/3072959.3073640

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A Deep Learning Approach for Generalized Speech Animation. *ACM Transactions on Graphics* 36, 4 (July 2017), 93:1–11. https://doi.org/10.1145/3072959.3073699

Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. 2012. Dynamic Units of Visual Speech. In *Symposium on Computer Animation (SCA)*. 275–284.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395. https://doi.org/10.1109/CVPR.2016.262

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2018. End-to-End Speech-Driven Facial Animation with Temporal GANs. In *British Machine Vision Conference (BMVC)*.

Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. 2018. X2Face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision (ECCV)*. https://doi.org/10.1007/978-3-030-01261-8_41

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *International Conference on Computer Vision (ICCV)*. 2868–2876. https://doi.org/10.1109/ICCV.2017.310

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *International Conference on Computer Vision (ICCV)*. 2242–2251. https://doi.org/10.1109/ICCV.2017.244

A APPENDIX

This appendix lists the used datasets in Table 4.

Table 4. List of datasets used in our results and comparisons as source and/or target videos. The language is given using ISO 639-2/B codes. Obama image courtesy of the White House (public domain).

Image	Name	Language	Style	#Frames
	David	eng	neutral + smile	3,232
	E.	ger	angry	8,628
	E.	ger	neutral	8,222
	E.	ger	squint	7,711
	F.	eng	neutral	8,020
	F.	ger	smile	8,333
	I.	ind	smile	8,782
	I.	eng	smile	10,138
	J.	eng	neutral	9,070
	J.	eng	sarcastic	8,977
	J.	eng	squint	8,048
	K.	eng	smile	9,972
	K.	ger	smile	11,745
	K.	eng	neutral	9,664
	M.	eng	smile	8,911
	M.	kan	smile	8,294
	Obama	eng	neutral	1,945
	Obama	eng	neutral	3,613
	Thomas	ger	neutral + smile	3,232