IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, USA, October 2023.

Attention-Based VR Facial Animation with Visual Mouth Camera Guidance for Immersive Telepresence Avatars

Andre Rochow, Max Schwarz, and Sven Behnke

***Abstract*— Facial animation in virtual reality environments is essential for applications that necessitate clear visibility of the user's face and the ability to convey emotional signals. In our scenario, we animate the face of an operator who controls a robotic Avatar system. The use of facial animation is particularly valuable when the perception of interacting with a specific individual, rather than just a robot, is intended. Purely keypoint-driven animation approaches struggle with the complexity of facial movements. We present a hybrid method that uses both keypoints and direct visual guidance from a mouth camera. Our method generalizes to unseen operators and requires only a quick enrolment step with capture of two short videos. Multiple source images are selected with the intention to cover different facial expressions. Given a mouth camera frame from the HMD, we dynamically construct the target keypoints and apply an attention mechanism to determine the importance of each source image. To resolve keypoint ambiguities and animate a broader range of mouth expressions, we propose to inject visual mouth camera information into the latent space. We enable training on large-scale speaking head datasets by simulating the mouth camera input with its perspective differences and facial deformations. Our method outperforms a baseline in quality, capability, and temporal consistency. In addition, we highlight how the facial animation contributed to our victory at the ANA Avatar XPRIZE Finals.**

I. Introduction

Facial animation is an important task in visual computing. A popular setting is face reenactment, where a source image and a driving image, which may be of different persons, are provided. The resulting image should have the appearance of the source image person, but the pose and facial expression of the driving image person. Generally, face reenactment methods are trained on speaking head datasets, such as Vox-Celeb [1]. At inference time, the objective is to utilize arbitrary driving videos to animate the source-image person.

A special case in virtual reality is VR facial animation, where the user wears a head-mounted display (HMD) and is, thus, not fully visible. Driving information has to be captured by sensors mounted on the headset and is typically incomplete. Limited and occluded information together with large perspective offsets makes VR facial animation exceptionally challenging. Furthermore, many HMDs cause deformations in even the visible areas which particularly limits mouth movements. The alignment problem between mouth camera images and images without the presence of an HMD is one of the biggest challenges for generating training samples.

Our system was developed for the ANA Avatar XPRIZE Challenge[1], where a previously unknown operator had to perform various tasks through our Avatar robot system (see Fig. 1). Our robotic system [2], [3] consists of an operator station with a VR headset and arm exoskeletons, as well as an avatar robot. We use a modified *Valve Index* HMD equipped with two infrared eye cameras and a mouth camera [4]. As visualized in Fig. 1, the operator's face is animated on a display that mirrors the operator head movement using an 6 DoF robotic arm. At the competition participants were judged not only on task performance, but also on immersion and the communication experience of a remote recipient. In particular, points were awarded when the operator was able to convey emotional cues to the recipient. Facial animation was thus a cornerstone of our strong performance at the challenge finals in November 2022, where our team NimbRo won the first prize.

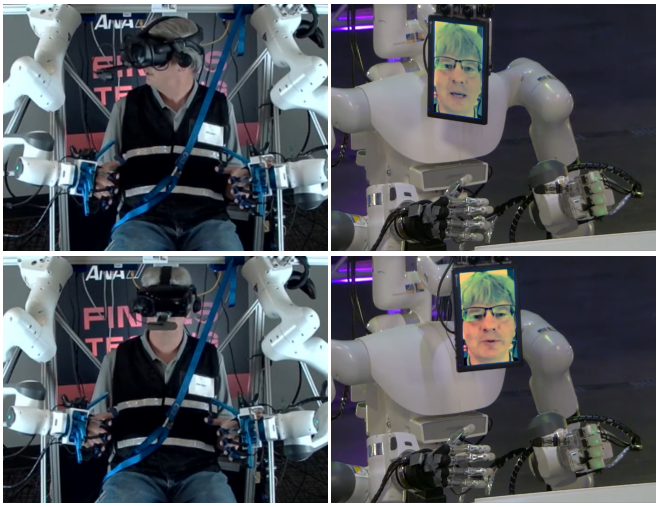


Fig. 1. Facial animation of an operator interacting with a recipient through the NimbRo Avatar system at the ANA Avatar XPRIZE Finals. Stills from our winning run[3]. Contrast was enhanced for easier viewing.

In our previous work [4], we formulated VR facial animation as a keypoint-driven face reenactment problem. This allowed us to train on large speaking head datasets and leverage knowledge obtained from many different appearances to animate unseen persons. In particular, we guided animation of the mouth area by dynamically retrieving a source image based on its keypoint similarity to the mouth camera image. Unfortunately, temporal inconsistencies occurred whenever changing the expression frame and the animation accuracy strongly depended on the image retrieval quality. Furthermore, keypoint ambiguities limit the range of possible expressions.

All authors are with the Autonomous Intelligent Systems group of University of Bonn, Germany; rochow@ais.uni-bonn.de

[1] https://www.xprize.org/prizes/avatar

[3] https://www.youtube.com/watch?v=OD2UbZNw9sQ

In this work, we propose an extension of [4] to address these limitations while preserving the ability to generalize to unseen persons. We propose to utilize multiple source images with an attention mechanism driven by the mouth camera, enabling our method to dynamically weight relevant features. Using the mouth camera video stream as the driving input reduces temporal inconsistencies, since the attention values are estimated by a continuous function that adapts to changes in the input.

We enhance the range of possible facial expressions and solve keypoint ambiguities by introducing a mouth camera guidance that directly utilizes visual mouth camera features. As discussed, the alignment problem makes it very challenging to generate suitable training data. We adress this issue by proposing an efficient way to keep training on large speaking head datasets for generalizability during inference and additionally annotate a few image pairs with similar mouth expressions in the mouth and face camera that we merge into the training process. As we demonstrate in a detailed evaluation and the supplementary video[4], our VR facial animation pipeline generates more accurate and more temporally consistent results than the baseline, with more movement in areas that are not associated with keypoints, such as the cheeks.

In addition to a real-time capable VR facial animation pipeline, our contributions include: (i) a source image attention mechanism that significantly improves temporal consistency and facial animation accuracy, (ii) an efficient way to leverage visual Mouth Camera information to resolve keypoint ambiguities and model a broader range of facial expressions, and (iii) emulation of mouth camera data, which allows training on available large-scale datasets.

II. Related Work

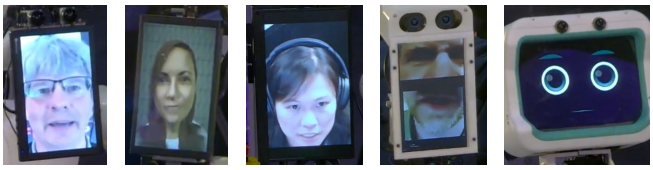
A. Face Reenactment

A task related to VR facial animation is face reenactment. Here, a driving frame which encodes the head pose and expressional information is to be visualized with the appearance given by a source image person. Often keypoints are used to represent the motion [5]–[7]. A motion network predicts a deformation grid to deform source images into a defined target motion. Siarohin *et al.* [6] propose to use image feature-based local affine transformations in the motion network that allow to model a larger family of transformations. Gafni *et al.* [8] propose to condition a dynamic NeRF with motion information extracted from driving images.

B. VR Facial Animation

In VR facial animation, the motion is encoded in eye camera images and a mouth camera image [4], [9], [10] or even in audio recordings [11]. Lombardi *et al.* [9] render a virtual avatar by utilizing a variational autoencoder (VAE) that can be conditioned with motion parameters obtained from the HMD. They train a second VAE on real and synthetic mouth camera images and map similar expressions of both domains to similar latent codes by manually controlling the latent variable that determines the domain. However, they do not handle facial deformations caused by the HMD explicitly. Wei *et al.* [10] generate synthetic ground truth data with an expression-preserving style transfer network, which maps between the mouth camera domain and the avatar domain. Richard *et al.* [11] bypass the alignment problem by omitting mouth camera images completely and using audio instead. They generate impressive results; however, the reduced amount of information significantly limits the expressivity. Especially when the user is silent, the animation task is ill-posed. Unfortunately, all these methods need a significant amount of data capture and operator-specific training, which makes them unsuitable for use-cases that require instant application, such as the ANA Avatar XPRIZE Competition.

[4]https://www.ais.uni-bonn.de/videos/IROS_2023_Rochow



a) Ours b) Mesh c) No VR d) Stitched e) Cartoon

Fig. 2. Types of facial animation at ANA Avatar XPRIZE finals. Examples from teams: a) NimbRo (first place), b) Pollen Robotics (second), c) Northeastern [12] (third), d) AVATRINA [13] (fourth), e) UNIST (sixth).

From ANA Avatar XPRIZE finals video footage we recognize five categories of face animation techniques used by participants (see Fig. 2). Out of the 12 teams selected for the two competition days, three teams had no VR headset. In this case, video streaming suffices for face display, but operator immersion is limited. Very similarly, one team displayed mouth camera footage directly, stitched together with previously captured footage of the operator's eyes. The rest of the teams used expression information from mouth trackers and/or audio to animate either 2D emoji drawings (three teams) or rendered 3D meshes, adapted to roughly match the operator's attributes like hair color and gender (four teams). Our team was the only one to produce a photorealistic animated face image.

III. Baseline

We select our avatar robot facial animation method developed in previous work [4] as a baseline, since it has the same input requirements and can thus be easily compared.

We give a brief overview of our previous work here. It is composed of 1) capturing and preprocessing, 2) image retrieval, 3) construction of the driving keypoints, 4) deforming, and 5) fusing and refining.

1) Capturing and Preprocessing: We capture two videos of the operator, with and without the HMD, respectively. The mouth camera only captures the lower facial area and the second (source image) video captures the complete frontal facing head of the operator. From the source video, we select an arbitrary source image which subsequently defines the operator appearance. Keypoints are extracted from selected frames showing different facial expressions. We differentiate between lower facial area keypoints $\mathcal{K}_{VR}$, which are also visible in the mouth camera, and facial keypoints $\mathcal{K}_F$ that

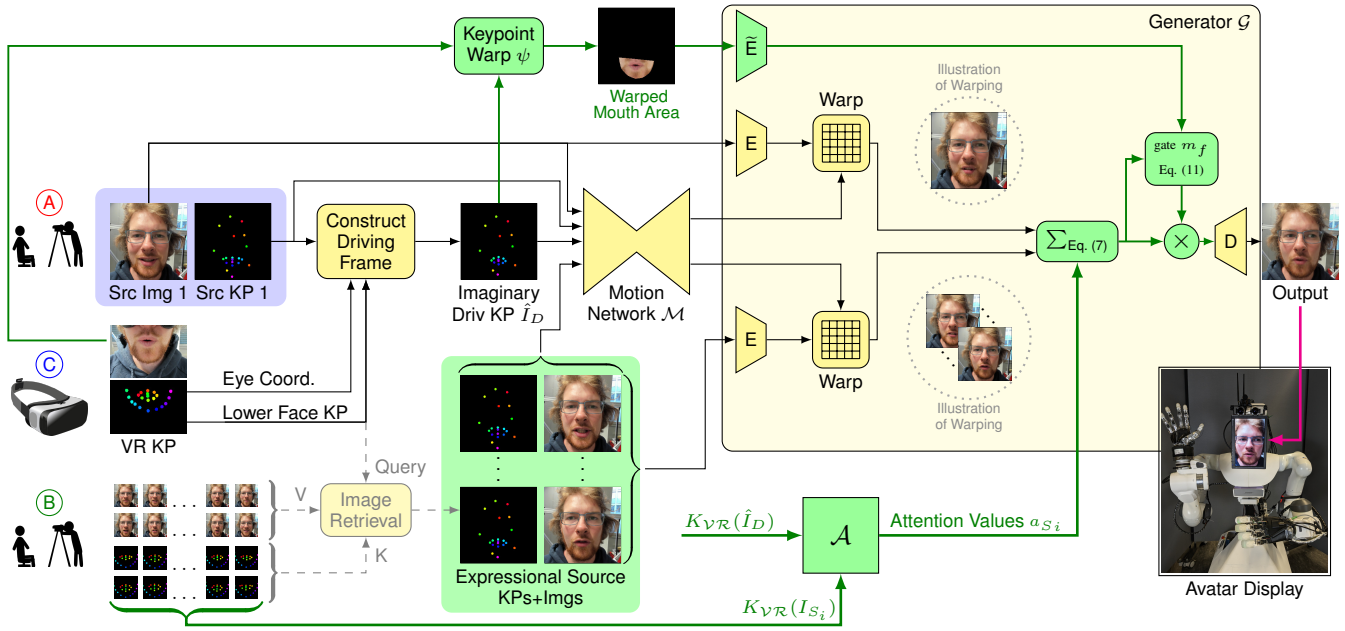


Fig. 3. Inference pipeline for VR Facial Animation. New components compared to our previous work [4] are highlighted in green. We select 4-5 still source images from a portrait video of the operator shot before the run as source images (A). The remaining frames are optionally used as a key-value storage of retrievable expression keypoints and corresponding images (B). The live keypoints measured inside and outside the VR headset (C) are then projected to the first source image frame, where they are optionally used to retrieve the closest expression image with keypoints from the storage. The keypoints of all source images including the retrieved one and a constructed set of driving keypoints then enter the motion network $\mathcal{M}$, which estimates a deformation grid that is used to warp the source images features, extracted by the generator-encoder network, to match the driving keypoints. The illustration of warping in $\mathcal{G}$ shows the deformation grid applied to the image instead of encoded features. The deformed features are aggregated over the source images in the lower facial area using a trainable attention mechanism $\mathcal{A}$. The mouth camera image from the HMD is warped into the lower facial area of the constructed driving keypoints and then encoded by a separate encoder network $\widetilde{E}$. An estimated mask m_f gates the aggregated deformed source features using the warped mouth camera features. The masked aggregated features are then decoded to produce the output.

determine the head pose including a keypoint kp_{eye} for the gaze direction. Given the set of mouth video keypoints $\{\mathcal{K}_{VR}(I_{M_0}), \mathcal{K}_{VR}(I_{M_1}), \ldots, \mathcal{K}_{VR}(I_{M_k})\}$ and source video keypoints $\{\mathcal{K}_{VR}(I_{S_0}), \mathcal{K}_{VR}(I_{S_1}), \ldots, \mathcal{K}_{VR}(I_{S_l})\}$, we define a keypoint mapping $\Pi_{S_i}(\mathcal{K}_{VR})$ that maps lower facial keypoints from the mouth camera into source image I_{S_i}. $\Pi(\cdot)$ corrects effects caused by the perspective change and deformations caused by the HMD's weight.

2) Image Retrieval: The image retrieval process searches the source video for a so-called *expression image* that has a similar mouth expression as the live mouth camera image. Given the projection $\Pi_{S_i}(\mathcal{K}_{\mathcal{VR}})$ of the mouth camera keypoints, we therefore retrieve the source video image I_{S_i} with the best matching keypoints. The expression image is then utilized to guide the animation process in the lower facial area.

3) Construction of the Driving Keypoints: The keypoints $\mathcal{K}(\hat{I}_D)$ of an imaginary driving frame that fully specify the facial target expression and pose are constructed here. The 3D keypoints that determine the head pose are simply copied from the source image $\mathcal{K}_F(I_S) = \mathcal{K}(\hat{I}_D)$ since we move the face display on the robot and thus do not require head movement in the output animation. The gaze keypoints are estimated by transferring eye tracking results from the eye cameras into a normalized gaze coordinate system in the source image. For a detailed explanation we refer to Rochow *et al.* [4]. The lower facial keypoints $\mathcal{K}_{VR}(\hat{I}_D)$ are generated by projecting the current mouth camera keypoints into the fixed source image. We thus define the imaginary driving keypoints

$$\mathcal{K}(\hat{I}_D) := \Pi_S(\,\mathcal{K}_{\mathcal{VR}}(I_M)\,) \oplus \rho(\,\mathcal{K}_{\mathcal{F}}(I_S), \hat{kp}_{eye}\,), \qquad (1)$$

where I_M is the mouth camera image, $\Pi_S(\cdot)$ maps each lower-face keypoint $kp_M^{(i)} \in \mathcal{K}_{\mathcal{VR}}(I_M)$ into the source image I_S, and $\rho(\cdot)$ replaces the eye keypoints detected in I_S with the modified values $\hat{kp}_{eye}$ in order to include the operator's current gaze direction and eye openness.

4) Deforming: The motion network $\mathcal{M}$ generates deformation grids $\mathcal{M}_{S\leftarrow D}$ and $\mathcal{M}_{E\leftarrow D}$ that are used to sample a deformation of the source and expression image into the imaginary driving keypoints. The motion network cannot generate new content, but it generates a good initialization for the refinement network.

5) Fusing and Refining: The refinement network $\mathcal{G}$ combines the deformed source image and the lower facial area of the deformed expression image. It generates a realistic output image with the appearance of the source image and the facial expressions as specified by the constructed imaginary driving keypoints.

IV. Method

Our proposed method is an extension of our previous work [4]. The basic modules and steps (see section III) remain, with important functionalities added into the pipeline and refinement network. This extended refinement network is called generator $\mathcal{G}$ (see Figs. 3 and 4).

A. Source Image Attention Mechanism

We address temporal inconsistencies, as occurring in our baseline method, by using more than two source images and introducing an attention mechanism that equips the network with the ability to decide on how much information it requires from each source view. The attention mechanism works in several stages. We distinguish between two types of input images, the appearance (or first) source image I_{S_1} and the expressional source images $I_{S_2}, I_{S_3}, \ldots, I_{S_n}$. The first source image conserves all the appearance information of the operator, whereas the expressional source images are used to generate more accurate animations, by presenting the network different variations of the lower facial area of an operator. Especially the mouth area has a lot of variations due to occlusions, disocclusions and a many degrees of freedom when speaking. Given the selected source images $I_{S_1}, I_{S_2}, \ldots I_{S_n}$ and the corresponding facial keypoints $\mathcal{K}(I_{S_1}), \mathcal{K}(I_{S_2}), \ldots, \mathcal{K}(I_{S_n})$ we extract all keypoint sequences that correspond to the lower facial area $\mathcal{K}_{VR}(I_{S_1}), \mathcal{K}_{VR}(I_{S_2}), \ldots, \mathcal{K}_{VR}(I_{S_n})$, which we call VR keypoints as they are also visible in the mouth camera of the HMD. For a sequence of VR keypoints $kp_j \in \mathcal{K}_{VR}(I_{S_i})$ of the source image I_{S_i}, we generate a distance tensor $\mathcal{D}_{S_i}$ with

$$\mathcal{D}_{S_i}^{k,l} = \frac{kp_k - kp_l}{\max \mathcal{D}_{S_i}} \in \mathbb{R}^2. \tag{2}$$

The distance tensor $\mathcal{D}_D$ is generated for the driving keypoints $\mathcal{K}_{VR}(I_D)$ analogously. We then estimate similarity vectors of the source distance tensors

$$\vec{x}_{S_i} = \vec{\mathcal{D}_{S_i}} W^S \in \mathbb{R}^{256} \tag{3}$$

and the driving distance tensor

$$\vec{x}_D = \vec{\mathcal{D}_D} W^D \in \mathbb{R}^{256}, \tag{4}$$

where $W^S, W^D \in \mathbb{R}^{d,256}$ are learned weight matrices and $\vec{\mathcal{D}}$ represents a flattened vector representation of a distance tensor. The similarity values are finally given by the scaled dot products

$$x_{S_i} = \frac{\vec{x}_{S_i} \vec{x}_D^{\,T}}{\sqrt{256}} \in \mathbb{R}, \tag{5}$$

which are fed into a softmax function to generate attention values $a_{S_i} \in \mathbb{R}$. These steps are summarized with $\mathcal{A}$ in Figs. 3 and 4.

Before we calculate the weighted sum we extract features $E_{S_i} = E(I_{S_1})$ of all source images I_{S_i}, using the generator encoder network E (see Fig. 3), and align them in the driving keypoints. This is achieved by deforming the features into the driving keypoints using the deformation grid $\mathcal{M}_{S_i \leftarrow D}$ estimated by the motion network. The deformation generates a roughly aligned feature representation

$${}^{D}E_{S_i} = \mathcal{M}_{S_i \leftarrow D}[\, E(I_{S_1})\,]. \tag{6}$$

The aggregated deformed source image features

$${}^{D}E_S = (1 - B_{LF})\, {}^{D}E_{S_1} + \sum_{i=1}^{n} a_{S_i} B_{LF}\, {}^{D}E_{S_i}, \tag{7}$$

are generated by a weighted sum in the lower facial area, where B_{LF} is a binary mask that crops out the lower facial area of the deformed source images features and a_{S_i} are the attention values.

B. Visual Mouth Camera Guidance

We address keypoint ambiguities by leveraging visual information from the current mouth camera image to guide the animation process. It is challenging to directly process the mouth camera image due to perspective changes and deformations caused by the HMD.

Our key idea for addressing this issue is to reuse the obtained lower facial (VR) keypoints from the mouth camera image $\mathcal{K}_{VR}(I_M)$ and its deformation-aware projection $\Pi_S(\mathcal{K}_{VR}(I_M))$ into the driving head pose. We first estimate a Delaunay triangulation and then use barycentric coordinates to sample the mouth camera image in the target keypoints. We define the mouth area keypoint warping

$$\psi(\, I_1, \mathcal{K}_{VR}(I_1), \mathcal{K}_{VR}(I_2)\,) \tag{8}$$

to be a function that samples the image I_1 with keypoints $\mathcal{K}_{VR}(I_1)$ in the keypoints $\mathcal{K}_{VR}(I_2)$ of image I_2. If we set $I_1 = I_M$ and $I_2 = I_D$ this gives us an approximation that accounts for the perspective change and the deformations caused by the HMD.

1) Mouth Camera Emulation during Training: Unfortunately, the alignment problem of mouth camera images and entire faces without an HMD makes it impossible to obtain perfect ground truth pairs for training. In our baseline approach [4], the information bottleneck posed by the VR keypoints enables training on large-scale speaking head datasets which helps generalization to unseen persons without finetuning.

To maintain this behavior and still provide visual mouth camera information, we propose a training-time data augmentation scheme. We add different types of camera noise [14], but also simulate imperfect transformation by performing a keypoint warping on the driving frame to itself ($I_1 = I_2 = I_D$) with noise added to the keypoints (see Fig. 4). The noise-augmented keypoint sequences are given by augmenting with (i) a normal distributed random scaling factor of the keypoint vector, (ii) a normal distributed random translation of the keypoint vector, and (iii) a normal distributed offset for each keypoint in the vector.

During training, the resulting keypoint warping function

$$\psi(\, \omega^I[\, I_D\,], \omega^K[\, \mathcal{K}_{VR}(I_D)\,], \omega^K[\, \mathcal{K}_{VR}(I_D)\,]\,) \tag{9}$$

therefore only utilizes I_D in combination with an image noise operator ω^I and a keypoint noise operator ω^K (see Fig. 4).

2) Gating Network: Additionally, we allow usage of the warped mouth area only through gated convolutions, which prevents direct information propagation. We feed the keypoint-warped representation of the mouth area into the mouth image encoder $\widetilde{E}$ (see generator in Figs. 3 and 4) that has a downsampling factor of four. This estimates the warped mouth area features

$$\widetilde{E}_M = \widetilde{E}(\, \psi(\, I, \mathcal{K}_{VR}(I), \mathcal{K}_{VR}(I_D)\,)\,), \tag{10}$$

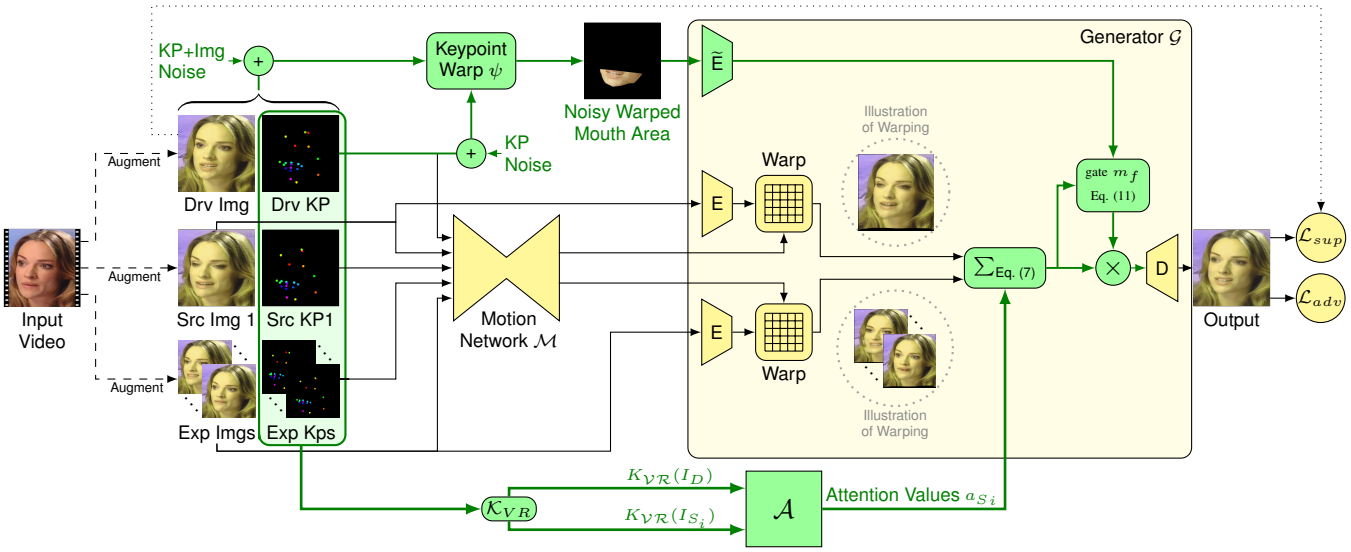


Fig. 4. Training the facial animation network from videos. New components compared to our previous work [4] are highlighted in green. The supervised training loss $\mathcal{L}_{sup}$ is minimized when the network reconstructs the driving image given the source images, source keypoints, noisy warped mouth area, and the driving keypoints. Furthermore, a keypoint-aware discriminator network judges the quality of the generated image ($\mathcal{L}_{adv}$). The source images are chosen randomly from the input video. During training, we simulate mouth camera guidance by utilizing the lower facial area of the driving image itself. We regularize the mouth camera guidance by injecting image noise and keypoint noise. This simulates different lighting and imperfect keypoint warping as present during inference when the real mouth camera image is utilized.

where $\psi(\cdot)$ performs the keypoint warping from image I into the driving VR keypoints $\mathcal{K}_{VR}(I_D)$.

For gating the aggregated deformed source features ${}^{D}E_S$, we concatenate the warped mouth area features $\widetilde{E}_M$ with ${}^{D}E_S$ and feed them through a small residual network with two layers to compute the gating weights (see Fig. 3). The resulting features in the main branch are therefore given by

$$f = \underbrace{\sigma(\ \phi[\widetilde{E}_M \oplus {}^{D}E_S]\ )}_{=:m_f} \odot {}^{D}E_S, \tag{11}$$

where $\odot$ is the elementwise multiplication, $\oplus$ is concatenation, $\sigma(\cdot)$ is the sigmoid function, and $\phi[\cdot]$ is the convolutional feature extraction of the small residual network.

Inducing visual mouth camera information implicitly through gating allows to mask out incorrect activations in the aggregated deformed source features ${}^{D}E_S$ (see Eq. (7)) while still being able to encode additional information without direct information propagation. This is especially beneficial when performing inter-operator animation (see Fig. 6) or generalizing from entire faces during training to mouth camera images during inference.

C. Training

The training pipeline is visualized in Fig. 4. All modules are trained end-to-end on the speaking head dataset VoxCeleb [1]. We train with perceptual loss and utilize a keypoint-aware discriminator network to generate adversarial losses, similar to Siarohin *et al.* [6]. Given a video, we randomly choose one driving frame and $n=5$ different source images, from which the last four are expressional source images. We extract facial keypoints and estimate a deformation grid of all source images into the driving keypoints using the motion network $\mathcal{M}$. Simultaneously, we estimate the attention values a_{S_i}. The source image features are then deformed and aggregated in the lower facial area using the attention values (see Eq. (7)). The features are conditioned with the keypoint warped mouth area (see Eq. (11)) and decoded to the output image.

We initialize the keypoint detector, motion network, generator-encoder, and generator-decoder with weights of our baseline. The new components (attention mechanism $\mathcal{A}$ and gating network) are trained from scratch. We found that the initialization with the baseline weights resulted in a very fast progress.

Finetuning with Imperfect VR Annotations: Unlike the baseline, our proposed method allows to explicitly train with mouth camera images. We therefore extend the datasets with some VR facial animation samples. We annotate such samples by manually searching for correspondences in the mouth camera and the face camera. The manual alignment is a very challenging task and often there is no perfect solution. Due to time limitations and efficiency reasons, we only annotate 13 different operators of our system and chose a fraction of training samples from such imperfect annotations. To prevent overfitting, we randomly scale, rotate and crop the facial images. Random cropping followed by rescaling to a quadratic image also changes face aspect ratio.

During finetuning, we select 6% of the training samples from the annotated VR datasets and 94 % from the VoxCeleb dataset, which gives similar importance to our annotated videos and videos from Vox-Celeb.

D. Inference

Preprocessing, which is explained in section III, remains equivalent to our baseline and takes approximately 15 minutes. We then select n=4 or n=5 fixed source images with different facial expressions. Given the current mouth camera image we optionally retrieve the best matching image

(see section III) which will be treated as an expressional source image. Following our baseline, we then construct the imaginary driving keypoints using the mapped mouth camera keypoints, the 3D head pose keypoints from the first source image, and the eye tracking results. The deformation, attention and refinement steps are equivalent to the training pipeline. During inference, the keypoint warping and gating step is always performed with the mouth camera image from the HMD. We therefore set $\widetilde{E}_M$ (see Eq. (10)) in Eq. (11) to

$$\widetilde{E}_M = \widetilde{E}(\ \psi(\ I_M, \mathcal{K}_{VR}(I_M), \Pi_S(\mathcal{K}_{VR}(I_M))\ )\ ), \qquad (12)$$

where I_M is the mouth camera image and $\Pi_S(\mathcal{K}_{VR}(I_M))$ are the mouth camera keypoints projected into the first source image (the VR keypoints of the imaginary driving frame).

E. Temporal Consistency

The baseline often struggles generating temporally consistent facial animations. The abrupt change of the expression frame induces the greatest negative influence.

Our proposed attention mechanism can be used in two different configurations. In the first configuration, all n=5 source frames are fixed, which minimizes the temporal inconsistencies as the continuous attention weight function changes smoothly with the mouth camera stream. The second configuration allows to retrieve the last expressional source image I_{S_5} during inference dynamically, which improves the output quality slightly (see Table I). To control the risk of temporal inconsistencies, we introduce a maximum attention value a_{max} for the retrieved images in the attention mechanism. This parameter allows us to control the tradeoff between quality and temporal consistency (see Table II). During testing, we set a_{max} operator-specifically but with a default value of 25%. In case the image retrieval does not perform well, the a_{max} value can be reduced.

Furthermore, the proposed visual mouth camera guidance reduces the network's dependency on the retrieved image which also contributes to the temporal consistency.

V. Experiments and Evaluation

We compare against our baseline method [4] which we used at the ANA Avatar XPRIZE Semifinals. A fair comparison to other methods [9]–[11] is not feasible as they perform per operator optimization with a significant amount of training, preprocessing, and data capturing. All reported qualitative and quantitative results are obtained with unseen persons.

A. Quantitative Results

To generate quantitative results, we utilize the annotated VR dataset. The mouth camera image is the input and the corresponding facial image will be the driving frame. We evaluate our method on five different persons. As our method is intended to improve the facial animation in the mouth region, we only measure the metrics Peak signal-to-noise ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [15] in the lower half of the face without background.

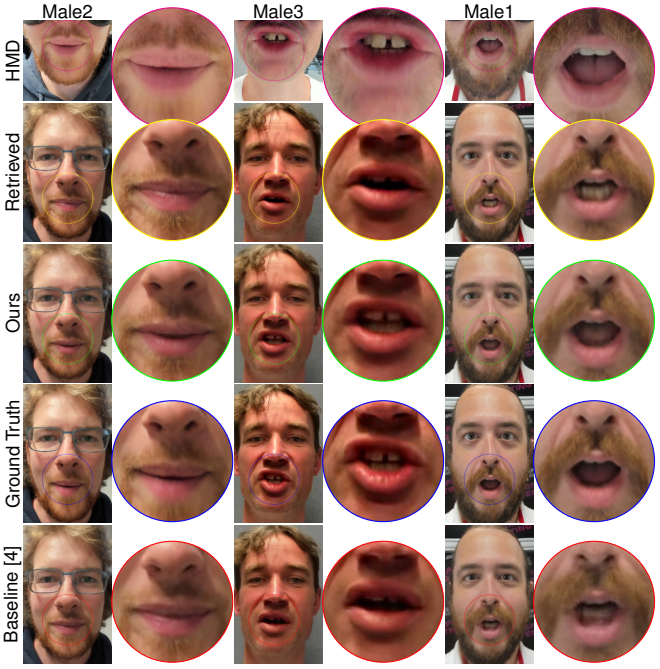


Fig. 5. Visual results of our quantitative analysis in Table I. For all examples the image retrieval (second row) was inaccurate, which led to poor results for the baseline [4] (bottom). Our method still generates good results.

1) Accuracy: Table I shows that all proposed model variants significantly outperform the baseline [4]. All of our ablations, besides Ours-Short and Ours-NF, in Table I are trained for 50 epochs with annotated VR samples as explained in section IV-C. The Ours-NF ablation, however, was never trained with VR samples. In this case, the missing regularizing influence results in overfitting after roughly five epochs, so we report the results at this training step. Interestingly, Ours-NF already outperforms the baseline in all metrics significantly. Ours-Short is only finetuned for 5 epochs which corresponds to just 4000 VR/Face image pairs that have been seen. This is already enough to generate similar results as obtained with 50 epochs VR finetuning.

Ours-10-Skip and Base-10-Skip represent ablations where only one out of ten images in the source video is retrievable. This results in a larger gap between the driving image and retrievable source images. When reducing retrievable images and thus the number of presented facial expressions by factor ten, quality is only influenced slightly (see mean metrics of Ours vs. Ours-10-Skip in Table I).

Our second method variation (Ours-5-Fix) further limits the number of different facial expressions presented to the network. It has only five fixed source images and therefore uses no image retrieval. The results indicate that image retrieval is not essential in our method for achieving good animations. Comparing all our method ablations shows that a_{max}=50% generates the highest accuracy, but results in reduced temporal consistency, compared to a_{max}=25% and Ours-5-Fix without image retrieval, as evaluated in Table II.

2) Temporal Consistency: In our proposed method and baseline [4] temporal inconsistencies mainly occur whenever a new expression frame is selected, which happens in roughly every second frame when speaking. Measuring temporal

TABLE I

ABLATION STUDY

Method	MEAN			Male1			Male2			Male3			Fem1			Fem2		
	psnr	ssim	lpips	psnr	ssim	lpips	psnr	ssim	lpips	psnr	ssim	lpips	psnr	ssim	lpips	psnr	ssim	lpips
Ours-50%	**28.83**	**.8603**	**.0357**	**29.27**	**.8642**	**.0365**	**28.66**	**.8610**	.0368	**28.59**	**.8439**	.0368	**29.87**	**.9028**	**.0233**	**27.74**	**.8298**	**.0451**
Ours	28.75	.8586	.0361	29.08	.8603	.0373	28.63	.8602	.0370	28.45	.8410	.0375	29.87	.9023	.0235	27.72	.8294	.0452
Ours-5-Fix	28.50	.8504	.0376	28.87	.8494	.0401	28.30	.8521	.0383	28.06	.8274	.0399	29.66	.8974	.0238	27.59	.8257	.0461
Ours-Short	28.28	.8550	.0376	28.64	.8486	.0437	28.45	.8589	**.0318**	28.19	.8436	**.0364**	28.69	.8963	.0246	27.42	.8275	.0515
Ours-NF	27.20	.8369	.0465	27.77	.8350	.0432	27.36	.8363	.0467	27.13	.8267	.0437	27.50	.8842	.0357	26.23	.8024	.0630
Base [4]	25.10	.7809	.0580	24.68	.7513	.0646	24.97	.7974	.0585	26.79	.7868	.0470	23.89	.8108	.0472	25.19	.7584	.0728
Ours-10-Skip	28.69	.8568	.0363	29.03	.8582	.0372	28.52	.8566	.0375	28.40	.8399	.0380	29.80	.9010	.0236	27.71	.8283	.0452
Base-10-Skip	24.96	.7758	.0596	24.61	.7469	.0653	24.91	.7902	.0589	26.69	.7858	.0481	23.78	.8074	.0500	24.83	.7486	.0756

NF: No finetuning on mouth camera images, *Short*: finetuning for a short time which leads to only 4000 VR images in the training batches, *10-Skip*: only one out of ten source video images retrievable, *Ours*: maximum image retrieval attention parameter $a_{max} = 25\%$, *Ours-50%*: $a_{max} = 50\%$, *Ours-5-Fix*: only five fixed source images without image retrieval.

consistency in animated facial images is a non-trivial task, especially when disoclusions and complex facial deformations occur. To reduce these effects, we use the motion network to deform the previous prediction into the current one. This allows comparison using perceptual similarity (LPIPS [15]), with the assumption that two consecutive frames exhibit only small expressional differences. Importantly, unintended discontinuous flicker effects lead to large errors in this metric. Note that the proposed measure does not necessarily correlate with accuracy.

In Table II, we report temporal inconsistency for four different persons from Table I, which are ordered with a descending image-retrieval quality from left to right. The best temporal consistency is obtained without image retrieval (Ours-5-Fix). When using image retrieval, the measured temporal consistency decreases with the maximum attention parameter a_{max} (see section IV-E). Together with Table I, this highlights the temporal consistency vs. accuracy tradeoff, which is controllable through a_{max}. However, compared to the baseline, all of our model variants perform much better, which is due to the baseline's strong dependence on the retrieved image. The discrepancy to our method gets larger with worsening image retrieval quality.

To increase temporal consistency, the baseline method (Base+TCF) explicitly minimizes this measuring scheme by recursively low-pass filtering the retrieved expression image using the deformations of the last expression frame and the last prediction, which is exactly what we measure. However, even if the image retrieval works fine, this comes with the cost of a reduced image quality in the lower facial area. Even though our ablations already achieve significantly better results than Base+TCF, we equip an additional ablation using $a_{max} = 50\%$ with the same recursive filtering scheme (Ours-50%+TCF) to allow a fairer comparison.

B. Qualitative Results

Qualitative results are shown in Figs. 5 to 7. Fig. 5 compares our method with ground truth and the baseline. It shows exemplary results of our quantitative evaluation in Table I. As can be seen, our results are much more accurate and closer to the ground truth. Unlike our method, the baseline fails whenever a bad expression image is retrieved.

Fig. 6. VR facial animation from mouth camera input to the appearance of a different operator. Mouth camera guidance resolves keypoint ambiguities and models a broader range of mouth expressions (note the lips which partly stick together).

TABLE II

TEMPORAL INCONSISTENCY

Method	Male2	Female1	Male3	Female2
Ours-5-Fix	**+0.0 %**	**+0.0 %**	**+0.0 %**	**+0.0 %**
Ours-25%	+6.2 %	+5.9 %	+11.5 %	+8.7 %
Ours-50%	+7.6 %	+8.3 %	+15.8 %	+16.8 %
Base [4]	+50.8 %	+86.2 %	+106.5 %	+151.9 %
Ours-50%+TCF	(+1.3 %)	(+1.3 %)	(+4.5 %)	(+4.4 %)
Base+TCF [4]	(+21.6 %)	(+33.5 %)	(+45.0 %)	(+88.1 %)

Values normalized to Ours-5-Fix. Lower is better. 25% and 50% indicate the a_{max} parameter, TCF means temporal consistency filtering [4]. Persons sorted by image retrieval quality (left: good).

Fig. 6 demonstrates very challenging mouth expressions obtained when mapping from the mouth camera input to a different person. This experiment shows that, unlike our baseline [4], the proposed mouth camera guidance allows to resolve keypoint ambiguities and properly displays very challenging facial expressions, such as lips which partly stick together.

Fig. 7 contains inference results compared with the baseline. In particular, we want to highlight that Ours-5-Fix produces almost the same results as our method configuration with image retrieval (Ours). The supplementary video (see Footnote 4) contains an animated comparison.

C. Throughput and Latency

We use pipelining techniques to enhance the throughput from 29 fps to 34 fps on an NVIDIA A6000 GPU with very

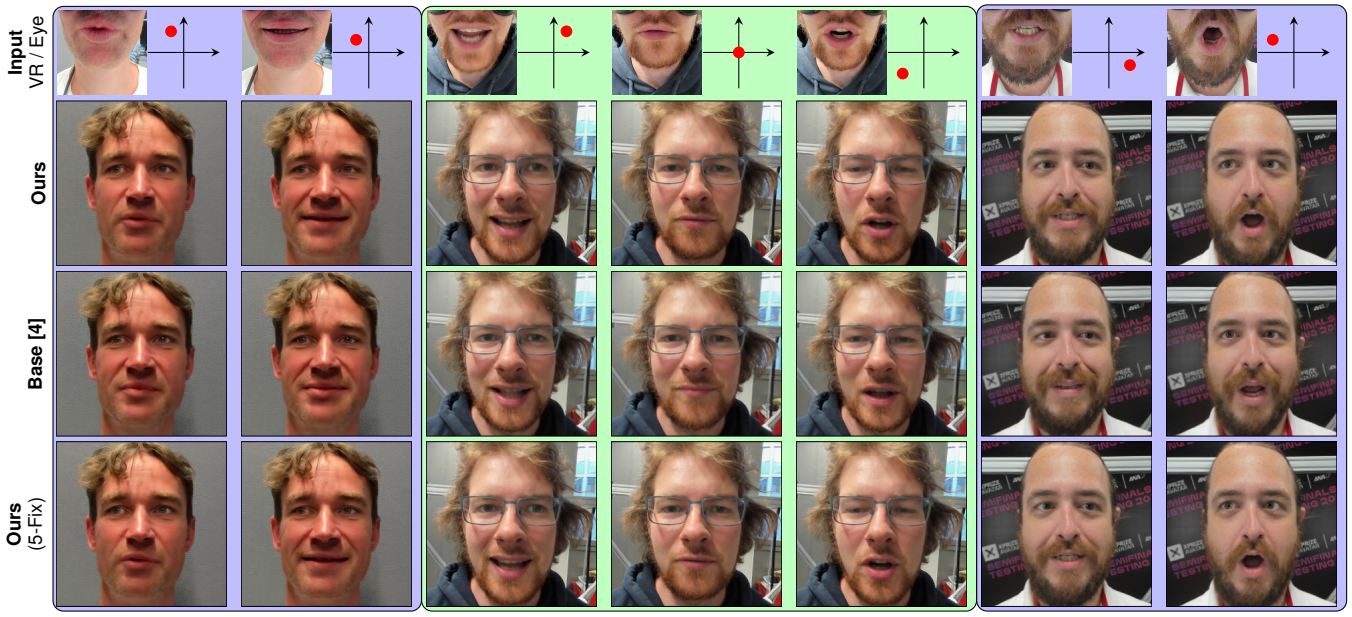


Fig. 7. Generated faces during inference, given mouth camera image and eye coordinates. See also the *supplementary video* for an animated comparison.

low latency (34 ms excluding and 51 ms including camera exposure time).

D. The ANA Avatar XPRIZE Finals

At the ANA Avatar XPRIZE competition finals in November 2022, our team and three different operators had to accomplish three test runs, of which the first one was a qualification run. The goal was to complete ten different tasks as fast as possible. For each completed task one point was awarded. Five additional points were awarded for usability and the ability to understand emotions and gestures. Especially for these, a facial animation was mandatory. Two tasks consisted of interacting with a human recipient. Our facial animation pipeline allowed seamless and immersive interaction between operator and recipient, which was rewarded with a full judge score on all three days. Overall, our Team NimbRo achieved a perfect score (15/15) with the fastest time in all three runs.

VI. Conclusion

We proposed a real-time capable VR facial animation approach that generalizes well to unseen operators and allows for modeling a broader range of facial expressions, compared to keypoint-driven approaches. We extended the baseline with a source image attention mechanism and developed a way to inject visual mouth image information into the animation pipeline without overfitting. These two extensions yield better accuracy and significantly improve temporal consistency which is important for smooth interaction. Our method still struggles in generating unusual expressions such as sticking out the tongue. Furthermore, movement in the upper face is still limited.

References

[1] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, 2020.

[2] M. Schwarz, C. Lenz, A. Rochow, M. Schreiber, and S. Behnke, "NimbRo Avatar: Interactive immersive telepresence with force-feedback telemanipulation," in *International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[3] M. Schwarz, C. Lenz, R. Memmesheimer, B. Pätzold, A. Rochow, M. Schreiber, and S. Behnke, "Robust immersive telepresence and mobile telemanipulation: NimbRo avatar wins ANA Avatar XPRIZE finals," *arXiv preprint arXiv:2303.03297*, 2023.

[4] A. Rochow, M. Schwarz, M. Schreiber, and S. Behnke, "VR facial animation for immersive telepresence avatars," in *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[5] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Neural Information Processing Systems (NeurIPS)*, 2019.

[7] R. Zhao, T. Wu, and G. Guo, "Sparse to dense motion transfer for face image animation," in *Int. Conf. Computer Vision (ICCV)*, 2021.

[8] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, "Dynamic neural radiance fields for monocular 4D facial avatar reconstruction," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh, "Deep appearance models for face rendering," *Trans. on Graphics (ToG)*, vol. 37, no. 4, 2018.

[10] S.-E. Wei, J. Saragih, T. Simon, A. W. Harley, S. Lombardi, M. Perdoch, A. Hypes, D. Wang, H. Badino, and Y. Sheikh, "VR facial animation via multiview image translation," *Trans. on Graphics (ToG)*, vol. 38, no. 4, 2019.

[11] A. Richard, C. Lea, S. Ma, J. Gall, F. De la Torre, and Y. Sheikh, "Audio-and gaze-driven facial animation of codec avatars," in *Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[12] R. Luo, C. Wang, E. Schwarm, C. Keil, E. Mendoza, P. Kaveti, S. Alt, H. Singh, T. Padir, and J. P. Whitney, "Towards robot avatars: Systems and methods for teleinteraction at Avatar XPRIZE semi-finals," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[13] J. M. Marques, N Patrick, Y. Zhu, N. Malhotra, and K. Hauser, "Commodity telepresence with the AvaTRINA nursebot in the ANA Avatar XPRIZE semifinals," in *RSS Workshop Towards Robot Avatars: Perspectives on the ANA Avatar XPRIZE Competition*, 2022.

[14] A. Carlson, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, "Modeling camera effects to improve visual learning from synthetic data," in *European Conference on Computer Vision (ECCV)*, 2018.

[15] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.