Fast Registration of Photorealistic Avatars for VR Facial Animation

Chaitanya Patel[1], Shaojie Bai[2], Te-Li Wang[2],
Jason Saragih[2], and Shih-En Wei[2]

[1] Stanford University, USA
[2] Meta Reality Labs, Pittsburgh, USA

https://chaitanya100100.github.io/FastRegistration/

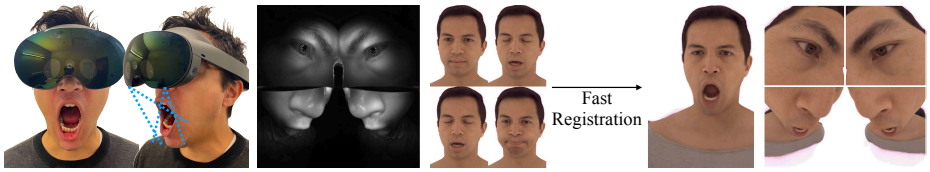

Fig. 1: On consumer VR headsets, **oblique mouth views and a large image domain gap hinder high quality face registration**. As shown, the subtle lip shapes and jaw movement are often hardly observed. Under this setting, our method is capable of efficiently and accurately registering facial expression and head pose of the photorealisitic avatars [7] of unseen identities.

Abstract. Virtual Reality (VR) bares promise of social interactions that can feel more immersive than other media. Key to this is the ability to accurately animate a personalized photorealistic avatar, and hence the acquisition of the labels for headset-mounted camera (HMC) images need to be efficient and accurate, *while* wearing a VR headset. This is challenging due to oblique camera views and differences in image modality. In this work, we first show that the domain gap between the avatar and HMC images is one of the primary sources of difficulty, where a transformer-based architecture achieves high accuracy on domain-consistent data, but degrades when the domain-gap is re-introduced. Building on this finding, we propose a system split into two parts: an iterative refinement module that takes in-domain inputs, and a generic avatar-guided image-to-image domain transfer module conditioned on current estimates. These two modules reinforce each other: domain transfer becomes easier when close-to-groundtruth examples are shown, and better domain-gap removal in turn improves the registration. Our system obviates the need for costly offline optimization, and produces online registration of higher quality than direct regression method. We validate the accuracy and efficiency of our approach through extensive experiments on a commodity headset, demonstrating significant improvements over these baselines.

Keywords: Face Registration · Virtual Reality · Image Style Transfer

1 Introduction

Photorealistic avatar creation has seen much progress in recent years. Driven by advances in neural representations and neural rendering [2, 22, 23, 30, 31], highly accurate representations of individuals can now be generated even from limited captures such as phone scans [7] or monocular videos [2, 15] while supporting real time rendering for interactive applications [26, 33]. Photorealistic quality is achieved by learning a universal prior model [7] on human appearance, which can be personalized to a novel user [7,15]. An emerging use case for such avatars is for enabling social interactions in Virtual Reality (VR). This application presents a particular problem where the user's face is typically occluded from the environment by the VR headset. As such, it relies on headset-mounted cameras (HMCs) to animate a user's avatar. While accurate results have been demonstrated, they have been restricted to *person-specific* cases, where correspondence pairs between the avatar and HMC images are obtained using additional elaborate capture rigs [33]. Highly accurate tracking in the more *general* case remains an open problem, due to the need of specializing a generic encoder to users' personalized avatars, *while* user is wearing a VR headset. Although fast adaptation methods are well studied [4, 9, 16], the unsolved challenge here is to obtain high quality image-label pair, especially under oblique camera viewing angles, time constraints, and the image domain difference between HMC images and avatar renderings.

In this work, we demonstrate that generic facial expression registration can be both accurate and efficient on unseen identities and challenging viewing angles. For this, we first demonstrate that accurate results are possible when the modalities of the headset-mounted cameras (typically infrared) and the user's avatar match, using a novel transformer-based network that iteratively refines expression estimation and head pose, solely from image features. Our method assumes no requirement on avatar to provide landmarks, which are not reliable under oblique HMC views. Building on this finding, we propose to learn a cross-identity style transfer function from the camera's domain to that of the avatar. The core challenge here lies in the high fidelity requirement of the style transfer due to the challenging viewpoints of the face presented by headset mounted cameras; even a few pixels error can lead to significant effects in the estimated avatar's expression. To resolve this, a key design of our method is to leverage an iterative expression and head pose estimation, as well as a style transfer module, which reinforce each other. On one hand, given a higher-quality style transfer module, the iterative refinement process gets increasingly easier. On the other hand, when a refined expression and pose estimation is closer to groundtruth, the style transfer network can easily reason locally using the input HMC images (conditioned on multiple *reference* avatar renderings) to remove the domain gap.

To demonstrate the efficacy of our approach, we perform experiments on a dataset of 208 identities, each captured in a multiview capture system [22] as well as a modified QuestPro headset [24], where the latter was used to provide ground truth correspondence between the driving cameras and the avatars. Compared to

direct regression method, our iterative construction shows significantly improved robustness against novel appearance variations in unseen identities.

In summary, the contribution of this work include:

- A demonstration that accurate and efficient generic face registration on a neural rendering face model is achievable under matching camera-avatar domains, without relying on 3D geometry.
- A generalizing style transfer network that precisely maintains facial expression on unseen identities, conditioned on photorealistic avatar renderings.
- Overall, a method to establish high-fidelity image-label pairs for unseen personalized avatars under time constraints and oblique viewing angles.

The remaining of the paper is structured as follows. In the next section, a literature review is presented. Then, in §3, we outline our method for generic facial expression estimation. In §4, we demonstrate the efficacy of our approach with extensive experiments. We conclude in §5 with a discussion of future work.

2 Related Work

2.1 VR Face Tracking

While face tracking is a long studied problem, tracking faces of VR users from head mounted cameras (HMCs) poses an unique challenge. The difficulty mainly comes from restrictions in camera placement and occlusion caused by the headset. Sensor images only afford oblique and partially overlapping views of facial parts. Previous work explored different ways to circumvent these difficulties. In [20], a camera was attached on a protruding mount to acquire a frontal view of the lower face, but with a non-ergonomic hardware design. In [32], the outside-in third-person view camera limits the range of a user's head pose. Both of these works rely on RGBD sensors to directly register the lower face with a geometry-only model. To reduce hardware requirements, [25] used a single RGB sensor for the lower face and performed direct regression of blendshape coefficients. The training dataset comprised of subjects performing a predefined set of expressions and sentences that had associated artist-generated blendshape coefficients. The inconsistencies between subject's performances with the blendshape-labeled animation limited animation fidelity.

A VR face tracking system on a consumer headset (Oculus Rift) with photoreaslitic avatars [22] was firstly presented in [33]. They introduced two novelties: (1) *The concept of a training- and tracking-headset*, where the former has a super-set of cameras of the latter. After training labels were obtained from the *training headset*, the auxiliary views from better positioned cameras can be discarded, and a regression model taking only *tracking headset*'s input was built. They also employed (2) *analysis-by-synthesis with differentiable rendering and style transfer* to precisely register parameterized photorealistic face models to HMC images, bridging the RGB-to-IR domain gap. The approach was extended in [30] via jointly learning the style-transfer and registration together, instead

of an independent CycleGAN-based module. Although highly accurate driving was achieved, both [33] and [30] relied on person-specific models, the registration process required hours to days of training, and required the *training headset* with auxiliary camera views to produce ground truth. As such, they cannot be used in a live setting where speed is required and only cameras on consumer headsets are available. In this work, we demonstrate that a system trained on a pre-registered dataset of multiple identities can generalize well to unseen identities' HMC captures within seconds. These efficiently generated image-label pairs can later be used to adapt a generic realtime expression regressor and make the animation more precise.

2.2 Image Style Transfer

The goal of image style transfer is to render an image in a target style domain provided by conditioning information, while retaining semantic and structural content from an input's content. Convolutional neural features started to be utilized [14] to encode content and style information. Pix2pix [18] learns conditional GANs along with L_1 image loss to encourage high-frequency sharpness, with an assumption of availability of paired ground truth. To alleviate the difficulty of acquiring paired images, CycleGAN [37] introduced the concept of cycle-consistency, but each model is only trained for a specific pair of domains, and suffers from semantic shifts between input and output. StarGAN [10] extends the concept to a fixed set of predefined domains. For more continuous control, many explored text conditioning [3] or images conditioning [1,8,11,21,34]. These settings usually have information imbalance between input and output space, where optimal output might not be unique. In this work, given a latent-space controlled face avatar [7], along with a ground-truth generation method [30], our style transfer problem can simply be directly supervised, with conditioning images rendered from the avatar to address the imbalance information problem.

2.3 Learning-based Iterative Face Registration

A common approach for high-precision face tracking involves a cascade of regressors that use image features extracted from increasingly registered geometry. One of the first methods to use this approach used simple linear models raw image pixels [29], which was extended by using SIFT features [36]. Later methods used more powerful regressors, such as binary trees [6, 19] and incorporated the 3D shape representation into the formulation. Efficiency could be achieved by binary features and linear models [28].

While these face tracking methods use current estimates of geometry to extract relevant features from images, similar cascade architectures have also been explored for general detection and registration. In those works, instead of *extracting* features using current estimates of geometry, the input data is augmented with *renderings* of the current estimate of geometry, which simplifies the backbone of the regressors in leveraging modern convolutional deep learning architectures. For example, Cascade Pose Regression [12] draws 2D Gaussians centered

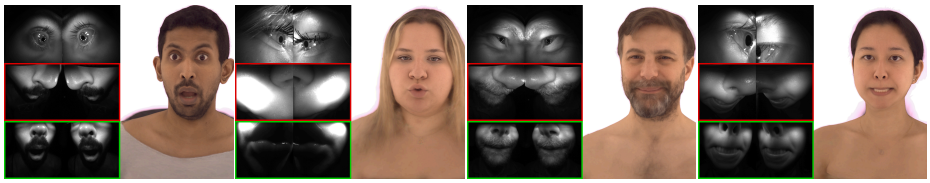

Fig. 2: Examples of HMC images and corresponding ground truth expression rendered on their avatars from the offline registration method [30], which utilizes augmented cameras with better frontal views (highlighted in green). In this work, we aim to efficiently register faces using cameras on consumer headsets, which only have oblique views (highlighted in red). In such views, information about subtle expressions (e.g., lip movements) are often covered by very few pixels or even not visible.

at the current estimates of body keypoints, which are concatenated with the original input, acting as a kind of soft attention map. Similar design in [5] was used for 3D heatmap prediction. Xia et al. [35] applied vision transformer [13] to face alignment with landmark queries. In this work, we demonstrate a transformer-based network that doesn't require any guidance from landmark to predict precise corrections of head pose and expression from multiview images.

3 Method

We aim to register the avatar face model presented in [7] to multi-view HMC images denoted $\boldsymbol{H} = \{H_c\}_{c \in C}$, where each camera view $H_c \in \mathbb{R}^{h \times w}$ is a monochrome infrared (IR) image and C is the set of available cameras on a consumer VR headset (in this work, we primarily focus on Meta's Quest Pro [24], see the supplementary material). They comprise a patchwork of non-overlapping views between each side of the upper and lower face. Some examples are shown in Fig. 2. Due to challenging camera angles and headset donning variations, it is difficult for the subtle facial expressions to be accurately recognized by machine learning models (e.g., see Fig. 7).

Setting. We denote the avatar's decoder model from [7] as $\mathcal{D}$. Following the same setting as in [7], given an input expression code $\boldsymbol{z} \in \mathbb{R}^{256}$, viewpoint $\boldsymbol{v} \in \mathbb{R}^6$, and identity information of the i^{th} subject, $\boldsymbol{I}^i$, the decoder is able to render this subject's avatar from the designated viewpoint by $R = \mathcal{D}(\boldsymbol{z}, \boldsymbol{v} | \boldsymbol{I}^i) \in \mathbb{R}^{h \times w \times 3}$. Specifically, when we use $\boldsymbol{v} = \boldsymbol{v}_c$; i.e., the viewpoint of a particular head-mounted camera (HMC), we'll obtain $R_c = \mathcal{D}(\boldsymbol{z}, \boldsymbol{v}_c | \boldsymbol{I}^i) \in \mathbb{R}^{h \times w \times 3}$, which has the same view as the corresponding $H_c \in \mathbb{R}^{h \times w}$, except the latter is monochromatic. Following [7], the identity information $\boldsymbol{I}^i$ for a specific identity i is provided as multi-scale untied bias maps to the decoder neural network. In this paper, we assume $\boldsymbol{I}^i$ is available for both training and testing identities, either from the lightstage or a phone scanning[3]; and that the calibrations of all head-mounted

[3] In this work we differentiate between unseen identities for avatar generation vs. unseen identities for HMC driving. We always assume an avatar for a new identity

cameras are known. We utilize the method in [30] to establish groundtruth HMC image-to-($\boldsymbol{z}$,$\boldsymbol{v}$) correspondences, which relies on an identity-specific costly optimization process and an augmented additional camera set, C', which provides enhanced visibility. The examples are highlighted in the green boxes in Fig. 2. Our goal in this work is to estimate the same optimal $\boldsymbol{z}$ and $\boldsymbol{v}$ for new identities leveraging the avatar model (i.e., registration), while using only the original camera set C, highlighted in red boxes in Fig. 2.

3.1 A Simplified Case: Matching Input Domain

Accurate VR face registration entails exact alignment between H_c and R_c for each head-mounted camera c. However, a vital challenge here is their enormous domain gap: $\boldsymbol{H} = \{H_c\}_{c \in C}$ are monochrome infrared images with nearfield lighting and strong shadows, while $\boldsymbol{R} = \{R_c\}_{c \in C}$ are renderings of an avatar built from uniformly lit colored images in the visible spectrum. [30,33] utilized a style transfer network to bridge this gap in a identity-specific setting. To simplify the problem in the generic, multi-identity case, we first ask the question: what performance is possible when there is no domain difference? To study this, we replace $\boldsymbol{H}$ with $\boldsymbol{R}_{gt} = \mathcal{D}(\boldsymbol{z}_{gt}, \boldsymbol{v}_{gt})$ obtained from the costly method in [30] with augmented cameras. $\boldsymbol{R}_{gt}$ can be seen as a perfectly style transferred result from $\boldsymbol{H}$ to the 3D avatar rendering space, that exactly retains expression. To predict $(\boldsymbol{z}_{gt}, \boldsymbol{v}_{gt})$ from $\boldsymbol{R}_{gt}$, a naïve way is to build a regression CNN which can be made extremely efficient such as MobileNetV3 [17]. Alternatively, given $\mathcal{D}$ is differentiable and the inputs are in the same domain, another straightforward approach is to optimize $(\boldsymbol{z}, \boldsymbol{v})$ to fit to $\boldsymbol{R}_{gt}$ using pixel-wise image losses. As we show in Table 1, the regression model is extremely lightweight but fails to generalize well; whereas this offline method (unsurprisingly) generates low error, at the cost of extremely long time to converge. Note that despite the simplification we make on the input domain difference (i.e., assuming access to $\boldsymbol{R}_{gt}$ rather than $\boldsymbol{H}$), the registration is still challenging due to the inherent oblique viewing angles, headset donning variations and the need to generalize to unseen identities.

In contrast, we argue that a carefully designed function that leverages avatar model (i.e., $\mathcal{D}$) information, which we denote as $\mathcal{F}_0(\cdot|\mathcal{D})$, achieves a good balance: (1) it is feed-forward (no optimization needed for unseen identities) so its speed can afford online usage; (2) it utilizes the renderings of $\mathcal{D}$ as a feedback to compare with input H_c and minimize misalignment. Before we describe $\mathcal{F}_0$ in § 3.3, we report the results of aforementioned methods under this simplified setting in Table 1.

Specifically, we show that $\mathcal{F}_0$ can achieve performance approaching that of offline registration [30]. In contrast, naïve direct regressions perform substantially worse, even with the augmented set of cameras. This highlights the importance of conditioning face registration learning with information about the target identity's avatar (in our case, $\mathcal{D}$). But importantly, when reverting back to the real

is already available through prior works, and evaluate the performance of expression estimation methods on unseen HMC images of that identity.

Table 1: Registration accuracy in a simplified setting. The errors are averaged across all frames in the test set. Augmented cameras means the use of camera set C' (which has better lower-face visibility) instead of C. Frontal Image L_1 describes expression prediction error, while rotation and translation errors describe the headpose prediction error. All methods are compared against groundtruth generated by the offline method [30] trained *with augmented cameras*. *Note that offline method below (colored in gray) is computed without augmented cameras, and is impractical due to the long convergence time.

	Aug. Cams	Input	Frontal Image L_1	Rot. Err. (deg.)	Trans. Err. (mm)	Speed
Offline [30]*	✗	$\boldsymbol{R}_{gt}$	0.784	0.594	0.257	∼1 day
Regression	✗	$\boldsymbol{R}_{gt}$	2.920	3.150	2.900	**7ms**
Regression	✓	$\boldsymbol{R}_{gt}$	2.902	3.031	3.090	**7ms**
Ours $\mathcal{F}_0$	✗	$\boldsymbol{R}_{gt}$	**1.652**	**0.660**	**0.618**	0.4sec
Ours $\mathcal{F}_0$	✓	$\boldsymbol{R}_{gt}$	**1.462**	**0.636**	**0.598**	0.4sec
Ours $\mathcal{F}_0$	✗	$\boldsymbol{H}$	2.851	1.249	1.068	0.4sec

problem, by replacing $\boldsymbol{R}_{gt}$ with $\boldsymbol{H}$, the performance of $\mathcal{F}_0$ also degrades significantly. This observation demonstrates the challenge posed by input domain gap difference, and motivates us to decouple style transfer problem from registration, as we describe next.

3.2 Overall Design

In light of the observation in §3.1, we propose to decouple the problem into the learning of two modules: an iterative refinement module, $\mathcal{F}$, and a style transfer module, $\mathcal{S}$. The goal of $\mathcal{F}$ is to produce an iterative update to the estimate expression $\boldsymbol{z}$ and headpose $\boldsymbol{v}$ of a given frame. However, as Table 1 shows, conditioning on avatar model $\mathcal{D}$ alone is not sufficient; good performance of such $\mathcal{F}$ relies critically on closing the gap between $\boldsymbol{H}$ and $\boldsymbol{R}_{gt}$. Therefore, module $\mathcal{F}$ shall rely on style transfer module $\mathcal{S}$ for closing this monochromatic domain gap. Specifically, in addition to raw HMC images $\boldsymbol{H}$, we also feed a style transferred version of them (denoted $\hat{\boldsymbol{R}}$), produced by $\mathcal{S}$, as input to $\mathcal{F}$. Intuitively, $\hat{\boldsymbol{R}}$ should then resemble avatar model $\mathcal{D}$'s renderings with the same facial expression as in $\boldsymbol{H}$. (And as Table 1 shows, if $\hat{\boldsymbol{R}} \approx \boldsymbol{R}_{gt}$, one can obtain really good registration.) Differing from the common style transfer setting, here the conditioning information that provides "style" to $\mathcal{S}$ is the entire personalized model $\mathcal{D}(\cdot|\boldsymbol{I}^i)$ itself. As such, we have the options of providing various conditioning images to $\mathcal{S}$ by choosing expression and viewpoints to render. Throughout experiments, we find that selecting frames with values closer to $(\boldsymbol{z}_{gt}, \boldsymbol{v}_{gt})$ improves the quality of $\mathcal{S}$'s style transfer output.

Therefore, a desirable mutual reinforcement is formed: the better $\mathcal{S}$ performs, the lower the errors of $\mathcal{F}$ are on face registration; in turn, the better $\mathcal{F}$ performs,

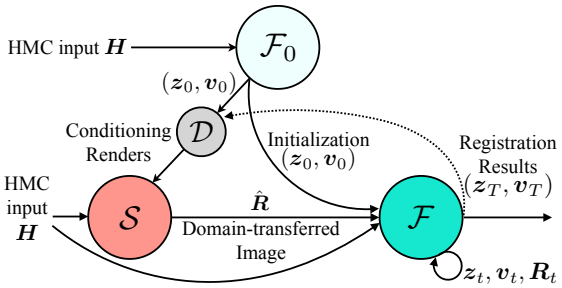

Fig. 3: Overview of the method. We decouple the problem into an avatar-conditioned image-to-image style transfer module $\mathcal{S}$ and a iterative refinement module $\mathcal{F}$. Module $\mathcal{F}_0$ initializes both modules by directly esimating on HMC input $\boldsymbol{H}$.

the closer rendered conditioning images will be to the ground truth, simplifying the problem for $\mathcal{S}$. An initialization $(\boldsymbol{z}_0, \boldsymbol{v}_0) = \mathcal{F}_0(\boldsymbol{H})$ for this reinforcement process can be provided by any model that directly works on monochromatic inputs $\boldsymbol{H}$. Fig. 3 illustrates the overall design of our system. In what follows, we will describe the design of each module.

3.3 Transformer-based Iterative Refinement Network

The role of the iterative refinement module, $\mathcal{F}$, is to predict the updated parameters $(\boldsymbol{z}_{t+1}, \boldsymbol{v}_{t+1})$ from input and current rendering:

$$[\boldsymbol{z}_{t+1}, \boldsymbol{v}_{t+1}] = \mathcal{F}\left(\boldsymbol{H}, \hat{\boldsymbol{R}}, \boldsymbol{R}_t\right), \quad \boldsymbol{R}_t = \mathcal{D}(\boldsymbol{z}_t, \boldsymbol{v}_t) \tag{1}$$

where $t \in [1, T]$ is number of steps and $\hat{\boldsymbol{R}} = \mathcal{S}(\boldsymbol{H})$ is the style-transferred result (see Fig. 4). $\mathcal{F}$ can reason about the misalignment between input $\boldsymbol{H}$ and current rendering $\mathcal{D}(\boldsymbol{z}_t, \boldsymbol{v}_t)$, with the aid of $\mathcal{S}(\boldsymbol{H})$ to bridge the domain gap.

In Fig. 4, we show the hybrid-transformer [13] based architecture of $\mathcal{F}$. For each view $c \in C$, a shared CNN encodes the alignment information between the current rendering $R_{t,c}$ and input images H_c along with style-transferred images $\hat{R}_c$ into a feature grid. After adding learnable grid positional encoding and camera-view embedding, the grid features concatenated with the current estimate $(\boldsymbol{z}_t, \boldsymbol{v}_t)$ and are flattened into a sequence of tokens. These tokens are processed by a transformer module with a learnable decoder query to output $(\Delta\boldsymbol{z}_t, \Delta\boldsymbol{v}_t)$, which is added to $(\boldsymbol{z}_t, \boldsymbol{v}_t)$ to yield the new estimate for the next iteration. We will show in §4.2 that this hybrid-transformer structure is a crucial design choice for achieving generalization across identities. The transformer layers help to fuse feature pyramid from multiple camera views while avoiding model size explosion or information bottleneck. Fig. 5 shows the progression of $\boldsymbol{R}_t$ over the steps. This iterative refinement module is trained to minimize:

$$\mathcal{L}_{\mathcal{F}} = \lambda_{\text{front}}\mathcal{L}_{\text{front}} + \lambda_{\text{hmc}}\mathcal{L}_{\text{hmc}}, \tag{2}$$

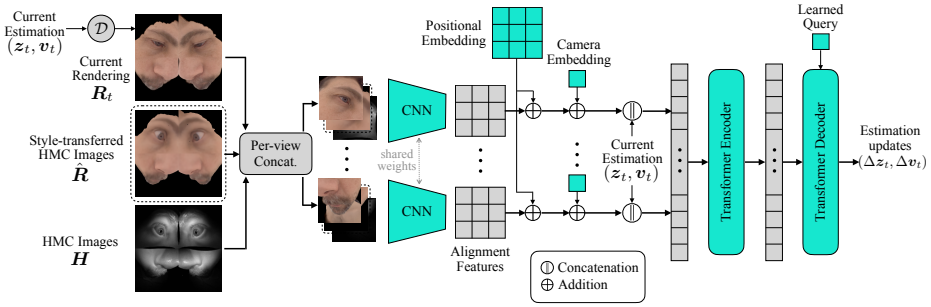

Fig. 4: Architecture of iterative refinement module $\mathcal{F}$

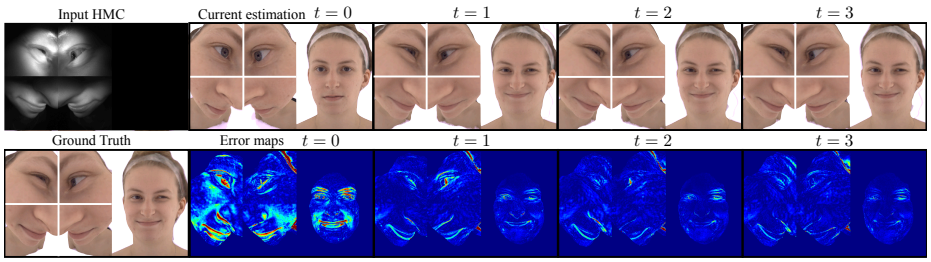

Fig. 5: Progression of iterative refinement in $\mathcal{F}$: we show intermediate results $\mathcal{D}(\boldsymbol{z}_t, \boldsymbol{v}_t)$ and corresponding error maps for each step t.

where

$$\mathcal{L}_{\text{hmc}} = \sum_{t=1}^{T} \sum_{c \in C} \|\mathcal{D}(\boldsymbol{z}_t, \boldsymbol{v}_{t,c} | \boldsymbol{I}^i) - \mathcal{D}(\boldsymbol{z}_{gt}, \boldsymbol{v}_{gt,c} | \boldsymbol{I}^i)\|_1$$

$$\mathcal{L}_{\text{front}} = \sum_{t=1}^{T} \|\mathcal{D}(\boldsymbol{z}_t, \boldsymbol{v}_{\text{front}} | \boldsymbol{I}^i) - \mathcal{D}(\boldsymbol{z}_{gt}, \boldsymbol{v}_{\text{front}} | \boldsymbol{I}^i)\|_1$$

Here, $\boldsymbol{v}_{\text{front}}$ is a predefined frontal view of the rendered avatar (see Fig. 5). While $\mathcal{L}_{\text{hmc}}$ encourages alignment between the predicted and groundtruth images from HMC views, $\mathcal{L}_{\text{front}}$ promotes an even reconstruction over the entire face to combat effects of oblique viewing angles in the HMC images.

While $\mathcal{F}_0$ could be any module that works on HMC images $\boldsymbol{H}$ for the purpose of providing $\{\boldsymbol{z}_0, \boldsymbol{v}_0\}$, for consistency, we simply set $\mathcal{F}_0$ to also be iterative refining, where the internal module is the same as $\mathcal{F}$, except without $\hat{\boldsymbol{R}}$ as input.

3.4 Avatar-conditioned Image-to-image Style Transfer

The goal of the style transfer module, $\mathcal{S}$, is to directly transform raw IR input images $\boldsymbol{H}$ into $\hat{\boldsymbol{R}}$ that resembles the avatar rendering $\boldsymbol{R}_{gt}$ of that original expression. Our setting differs from the methods in the literature in that our

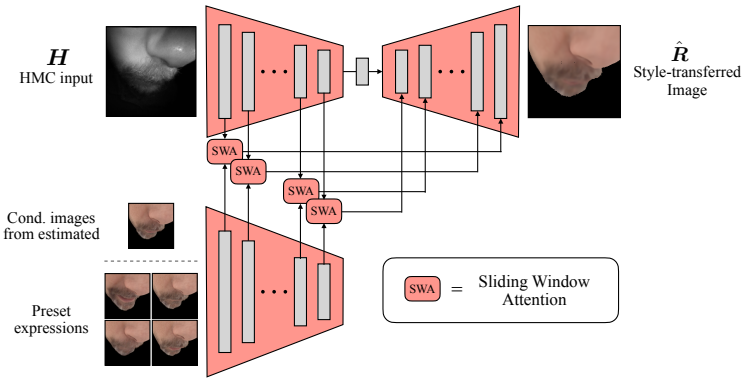

Fig. 6: Architecture of style transfer module $\mathcal{S}$

style-transferred images need to recover identity-specific details including skin-tone, freckles, etc., that are largely missing in the IR domain; meanwhile, the illumination differences and oblique view angle across identities imply any minor changes in the inputs could map to a bigger change in the expression. These issues make the style transfer problem ill-posed without highly detailed conditioning.

To this end, we design a novel style transfer architecture that utilizes the prior registration estimation given by $\mathcal{F}_0$. Specifically, we can utilize $\mathcal{F}_0$ that was trained directly on monochrome images $\boldsymbol{H}$, to obtain an estimate of $(\boldsymbol{z}_0, \boldsymbol{v}_0)$ for the current frame. Additionally, we choose M *reference conditioning expressions*: $(\boldsymbol{z}_{k_1}, ..., \boldsymbol{z}_{k_M})$ to cover a range of reference expressions; e.g., mouth open, squinting eyes, closed eyes, etc., which we find to significantly help mitigate ambiguities in style-transferring extreme expressions (we show examples of these conditioning reference expressions in the supplementary material). Formally, given the current frame HMC image $\boldsymbol{H}$, we compute

$$\hat{\boldsymbol{R}} = \mathcal{S}\left(\boldsymbol{H}, (\boldsymbol{z}_0, \boldsymbol{z}_{k_1}, ..., \boldsymbol{z}_{k_M}), \boldsymbol{v}_0\right). \tag{3}$$

With a better estimation of $(\boldsymbol{z}_0, \boldsymbol{v}_0)$ provided by $\mathcal{F}_0$, these conditioning images become closer to ground truth, thereby simplifying the style transfer learning task of $\mathcal{S}$.

Fig. 6 shows the UNet-based architecture of $\mathcal{S}$. Given an estimate of $(\boldsymbol{z}_0, \boldsymbol{v}_0)$, conditioning images are generated from the same estimate and M other key expressions, concatenated channel-wise and encoded by a U-Net encoder. Input HMC image is encoded by a separate U-Net encoder. Sliding window based attention [27] modules are used to fuse input features and conditioning features to compensate for the misalignment between them. These fused features are provided as the skip connection in the U-Net decoder to output style-transferred image $\hat{\boldsymbol{R}}$. This style transfer module is trained with a simple image L_1 loss:

$$\mathcal{L}_{\mathcal{S}} = \|\hat{\boldsymbol{R}} - \boldsymbol{R}_{gt}\|_1. \tag{4}$$

Table 2: Comparison of our approach (with style transfer *and* iterative refinement) with direct regression and offline methods. The errors are the averages of all frames in the test set. Augmented view means the use of camera set C' instead of C. All methods are comparing against groundtruth generated by the offline method [30] trained *with augmented cameras*. *Note that offline methods below (colored in gray) are computed without augmented cameras, and are impractical due to the long convergence time.

	Aug. cams	Input	Frontal Image L_1	Rot. Err. (deg.)	Trans. Err. (mm)	Speed
Offline [30]*	✗	$\boldsymbol{H}$	1.713	2.400	2.512	∼1 day
Offline [30]*	✗	$\boldsymbol{R}_{gt}$	0.784	0.594	0.257	∼1 day
Regression	✗	$\boldsymbol{H}$	2.956	2.850	2.802	**7ms**
Regression	✗	$\boldsymbol{R}_{gt}$	2.920	3.150	2.900	**7ms**
Regression	✓	$\boldsymbol{H}$	2.967	2.806	2.953	**7ms**
Regression	✓	$\boldsymbol{R}_{gt}$	2.902	3.031	3.090	**7ms**
Ours ($\mathcal{F}$+$\mathcal{S}$)	✗	$\boldsymbol{H}$	**2.655**	**0.947**	**0.886**	0.4s
Ours ($\mathcal{F}$+$\mathcal{S}$)	✓	$\boldsymbol{H}$	**2.399**	**0.917**	**0.845**	0.4s

4 Experiments

We perform experiments on a dataset of 208 identities ($1M$ frames in total), each captured in a lightstage [22] as well as a modified Quest-Pro headset [24] with augmented camera views. The avatars are generated for all identities with a unified latent expression space using the method from [7]. We utilize the extensive offline registration pipeline in [30](with augmented camera set C') to generate high-quality labels. We held out 26 identities as validation set. We use $T = 3$ refinement iterations during training and $M = 4$ key expressions to provide conditioning images for style transfer, which is operating at 192×192 resolution. See the supplementary material for more details on model architecture and training.

4.1 Comparison with Baselines

As discussed, there are two obvious types of methods to compare for general face registration: (1) the same **offline registration** method in [30], but only using the camera set C, performed individually on each validation identity's headset data. Since the training here is only across frames from that identity, it has the advantage to overfit on the same identity. However, it also limits the amount of prior knowledge it can leverage from other identities' images. Its performance anchors the challenge from camera angles, if computing time is not limited. (2) **Direct regression**: using the same set of ground truth labels, we train a MobileNetV3 [17] to directly regress HMC images to expression codes $\boldsymbol{z}$. This method represents an online model that could be used in a realtime system where the use of $\mathcal{D}$ is prohibited.

Table 2 summarizes the comparison. The offline method achieves good average frontal image loss. Albeit its high precision, it has common failure modes in

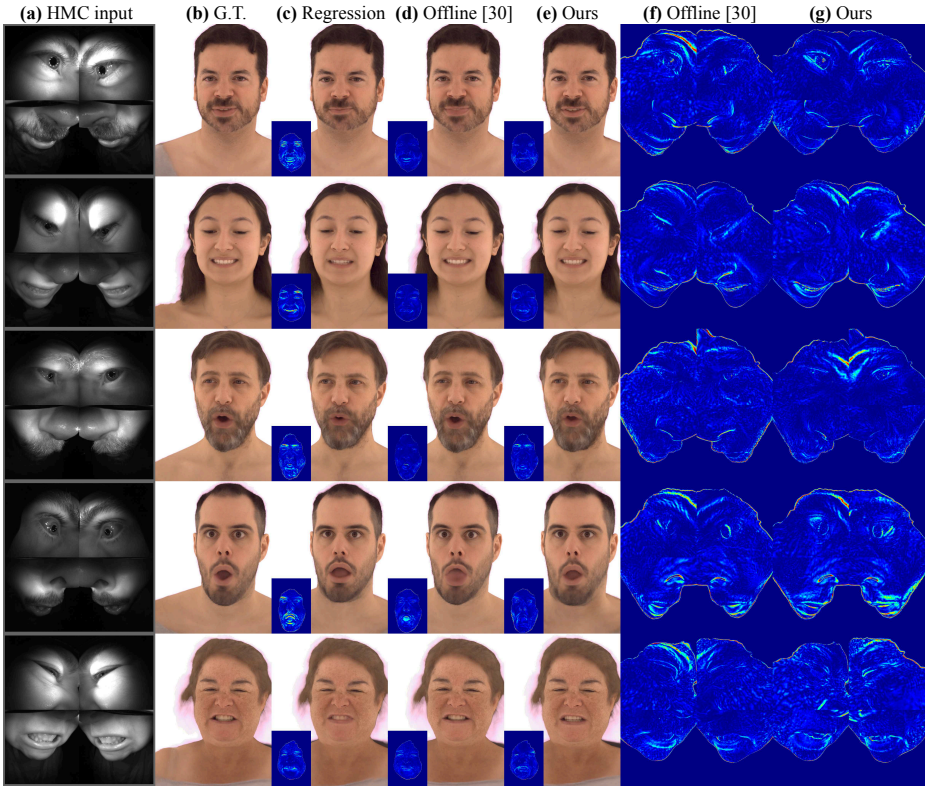

Fig. 7: Qualitative Results: we compare different methods by evaluating **(b,c,d,e)** frontal rendering (with error maps), and **(f,g)** error maps in HMC viewpoints. More examples are provided in the supplementary material.

lower jaw and inner mouth, where the observation is poor, as shown in Fig. 7. In comparison, our method could leverage the learning from cross-identity dataset, producing a more uniformly distributed error. The offline method also suffers from worse head pose estimation because its co-optimized style transfer could compensate small errors in oblique viewing angle. Our method is much faster due to its feed-forward design, enabling online generation of accurate labels.

On the other hand, the direct regression method generalizes poorly to novel identities, leading to worse performance on average. It also yields inferior results in estimating head poses. The head pose is defined as the relative 3D transformation from a reference camera to the avatar center which is not consistent across identities and not observable from HMC images. Since the regression baseline is not conditioned on avatar, there is no information to predict head poses accurately. We also provide relaxed conditions (e.g. $\boldsymbol{R}_{gt}$ as input, or using augmented cameras), and interestingly it fails to improve, while our method can leverage these conditions significantly. Our method's high accuracy, especially in the lip region as depicted in the supplementary video, captures nuanced

Table 3: Ablation on the design of $\mathcal{F}$. All methods use $\boldsymbol{R}_{gt}$ as inputs and without augmented cameras.

	Frontal Image L_1	Rot. Err. (deg.)	Trans. Err. (mm)
Ours $\mathcal{F}_0$	1.652	0.660	0.618
w/o transformer	2.533	2.335	2.023
w/o grid features	2.786	2.818	3.081
w/o transformer & w/o grid features	3.645	5.090	5.839

Table 4: Ablation on the design of $\mathcal{S}$.

	Image L_1 Error
Ours $\mathcal{S}$	2.55
w/o SWA	2.82
w/o key cond. expressions	2.75
w/o $\mathcal{F}_0$	2.99

facial expressions more effectively. These high-quality, quickly generated labels can be employed to adapt realtime regressors, thereby enhancing the immersive experience in virtual reality.

4.2 Ablation Studies

Iterative Refinement Module $\mathcal{F}$. Key to our design of $\mathcal{F}$ is the application of transformer on the grid of features from all camera views. We validate this design by comparing the performance of $\mathcal{F}_0(\boldsymbol{R}_{gt})$ against the following settings:

- **w/o transformer**, where we replace the transformer with an MLP. In this case, the grid features from all four camera views are simply concatenated and processed by an MLP. This trivial concatenation results in a 2x increase in the number of trainable parameters and subpar generalization.
- **w/o grid features**, where we average pool grid features to get a single feature for each camera view and use the same transformer design to process $|C|$ tokens.
- **w/o transformer & w/o grid features**, where we use an MLP to process the concatenation of pooled features from all camera views.

Results are shown in Table 3. We can see that processing grid features using transformer results in better generalization while requiring fewer parameters compared to using an MLP with trivial concatenation. Pooling grid features is also detrimental because it undermines minor variations in input pixels which are important in the oblique viewing angles of headset cameras. Transformer operating on grid tokens can effectively preserve fine-grained information and extract subtle expression details.

Style Transfer Module $\mathcal{S}$. We validate our design of $\mathcal{S}$ by comparing it with the following baselines:

- **w/o SWA**, where we simply concatenate the features of input branch with the features of conditioning branch at each layer.

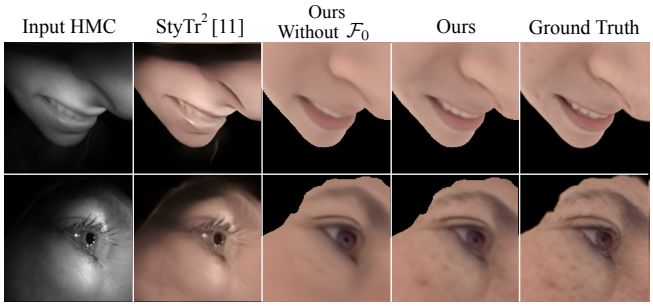

Fig. 8: Ablation on style transfer results. We compare our results with a generic style transfer method and baseline methods without the estimates provided by $\mathcal{F}_0$.

- **w/o key conditioning expressions**, where only the conditioning corresponding to the current estimate $(\boldsymbol{z}_0, \boldsymbol{v}_0)$ is used.
- **w/o $\mathcal{F}_0$**, where conditioning is comprised only of the four key expressions rendered using the average viewpoint per-camera, $\boldsymbol{v}_{\text{mean}}$.

Table 4 shows the average L_1 error between the foreground pixels of the groundtruth image and the predicted style transferred image. The larger error of style-transfer without $\mathcal{F}_0$ validates our design that a better style transfer can be achieved by providing conditioning closer to the groundtruth $(\boldsymbol{z}_{gt}, \boldsymbol{v}_{gt})$. When not incorporating SWA or key conditioning expressions, the model performs poorly when the estimates $\boldsymbol{v}_0$ and $\boldsymbol{z}_0$ are suboptimal respectively, resulting in higher error.

Fig. 8 shows qualitative results of style transfer. Here, we also show the result of StyTr^2 [11] - one of the recent style transfer methods that leverages the power of vision transformers [13] with large datasets. Despite using the groundtruth $\boldsymbol{R}_{gt}$ as the style image, it struggles to accurately fill in shadows and fine facial features that are not visible in the input HMC image. Although 'Without $\mathcal{F}_0$' produces better style transfer than StyTr^2 [11], it smooths out high-frequency details including freckles, teeth, soft-tissue deformations near eyes and nose. These high-frequency details are crucial for animating subtle expressions. Our style transfer model $\mathcal{S}$ is able to retain such details by leveraging the estimate provided by $\mathcal{F}_0$. See the supplementary material for more results.

5 Conclusion and Future Work

In this paper, we present a generic and feed-forward method for efficient registration of photorealistic 3D avatars on monochromatic images with oblique viewing angles. We show that closing the domain gap between avatar's rendering and headset images is a key to achieve high registration quality. Motivated by this, we decompose the problem into two modules, style transfer and iterative refinement, and present a system where one reinforces the other. Extensive

experiments on real capture data show that our system achieves superior registration quality than direct regression methods and can afford online usage. Our method provides a viable path for efficiently generating high quality label of neural rendering avatars on the fly, so that the downstream real-time model can adapt to achieve higher accuracy. This will enable the user to have photorealistic telepresence in VR without extensive data capture. In the future, extensions of our method could be done for general registration of neural rendering models on out-of-domain multi-view images, such as (non-VR) face registration, body tracking, and 3D pose estimation.

References

1. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: Artflow: Unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 862–871 (2021)
2. Apple Inc.: Apple Vision Pro. https://www.apple.com/apple-vision-pro/ (2024)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18392–18402 (June 2023)
4. Browatzki, B., Wallraven, C.: 3fabrec: Fast few-shot face alignment by reconstruction. In: CVPR (2020)
5. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
6. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. **33**(4) (jul 2014). https://doi.org/10.1145/2601097.2601204, https://doi.org/10.1145/2601097.2601204
7. Cao, C., Simon, T., Kim, J.K., Schwartz, G., Zollhoefer, M., Saito, S.S., Lombardi, S., Wei, S.E., Belko, D., Yu, S.I., Sheikh, Y., Saragih, J.: Authentic volumetric avatars from a phone scan. ACM Trans. Graph. **41**(4) (jul 2022). https://doi.org/10.1145/3528223.3530143, https://doi.org/10.1145/3528223.3530143
8. Chen, H., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D., et al.: Artistic style transfer with internal-external learning and contrastive learning. Advances in Neural Information Processing Systems **34**, 26561–26573 (2021)
9. Chen, L., Cao, C., la Torre, F.D., Saragih, J., Xu, C., Sheikh, Y.: High-fidelity face tracking for ar/vr via deep lighting adaptation (2021)
10. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
11. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr^2: Image style transfer with transformers. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
12. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1078–1085 (2010). https://doi.org/10.1109/CVPR.2010.5540094

13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2020), https://api.semanticscholar.org/CorpusID:225039882
14. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
15. Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L., Nießner, M.: Mononphm: Dynamic head reconstruction from monocular videos. arXiv preprint arXiv:2312.06740 (2023)
16. Guo, J., Zhu, X., Zhao, C., Cao, D., Lei, Z., Li, S.Z.: Learning meta face recognition in unseen domains. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6162–6171 (2020). https://doi.org/10.1109/CVPR42600.2020.00620
17. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
19. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
20. Li, H., Trutoiu, L., Olszewski, K., Wei, L., Trutna, T., Hsieh, P.L., Nicholls, A., Ma, C.: Facial performance sensing head-mounted display. ACM Transactions on Graphics (TOG) **34**(4), 47:1–47:9 (Jul 2015)
21. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6649–6658 (2021)
22. Lombardi, S., Saragih, J., Simon, T., Sheikh, Y.: Deep appearance models for face rendering. ACM Trans. Graph. **37**(4), 68:1–68:13 (Jul 2018)
23. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019)
24. Meta Inc.: Meta Quest Pro: Premium Mixed Reality. https://www.meta.com/ie/quest/quest-pro/ (2023)
25. Olszewski, K., Lim, J.J., Saito, S., Li, H.: High-fidelity facial and speech animation for vr hmds. ACM Transactions on Graphics (TOG) **35**(6), 1–14 (Nov 2016)
26. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. arXiv preprint arXiv:2312.02069 (2023)
27. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/3416a75f4cea9109507cacd8e2f2aefc-Paper.pdf
28. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1685–1692 (2014). https://doi.org/10.1109/CVPR.2014.218

29. Saragih, J., Goecke, R.: Iterative error bound minimisation for aam alignment. In: Proceedings of the 18th International Conference on Pattern Recognition - Volume 02. p. 1196–1195. ICPR '06, IEEE Computer Society, USA (2006). https://doi.org/10.1109/ICPR.2006.730, https://doi.org/10.1109/ICPR.2006.730
30. Schwartz, G., Wei, S.E., Wang, T.L., Lombardi, S., Simon, T., Saragih, J., Sheikh, Y.: The eyes have it: An integrated eye and face model for photorealistic facial animation. ACM Trans. Graph. **39**(4) (aug 2020). https://doi.org/10.1145/3386569.3392493, https://doi.org/10.1145/3386569.3392493
31. Shysheya, A., Zakharov, E., Aliev, K.A., Bashirov, R., Burkov, E., Iskakov, K., Ivakhnenko, A., Malkov, Y., Pasechnik, I., Ulyanov, D., Vakhitov, A., Lempitsky, V.S.: Textured neural avatars. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2382–2392 (2019), https://api.semanticscholar.org/CorpusID:160009798
32. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Niessner, M.: Facevr: Real-time gaze-aware facial reenactment in virtual reality. ACM Transactions on Graphics (TOG) **37**(2), 25:1–25:15 (Jun 2018)
33. Wei, S.E., Saragih, J., Simon, T., Harley, A.W., Lombardi, S., Perdoch, M., Hypes, A., Wang, D., Badino, H., Sheikh, Y.: Vr facial animation via multiview image translation. ACM Trans. Graph. **38**(4) (jul 2019). https://doi.org/10.1145/3306346.3323030, https://doi.org/10.1145/3306346.3323030
34. Wu, X., Hu, Z., Sheng, L., Xu, D.: Styleformer: Real-time arbitrary style transfer via parametric style composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14618–14627 (2021)
35. Xia, J., Qu, W., Huang, W., Zhang, J., Wang, X., Xu, M.: Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4042–4051 (2022). https://doi.org/10.1109/CVPR52688.2022.00402
36. Xiong, X., la Torre, F.D.: Supervised descent method and its applications to face alignment. 2013 IEEE Conference on Computer Vision and Pattern Recognition pp. 532–539 (2013), https://api.semanticscholar.org/CorpusID:608055
37. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)