Learning to Deblur and Rotate Motion-Blurred Faces

Givi Meishvili[1]
givi.meishvili@inf.unibe.ch

Attila Szabó[3]
attila.szabo@huawei.com

Simon Jenni[2]
jenni@adobe.com

Paolo Favaro[1]
paolo.favaro@inf.unibe.ch

[1] University of Bern, Switzerland

[2] Adobe Research

[3] Huawei, Noah's Ark Lab
(work was done before joining)

Abstract

We propose a solution to the novel task of rendering sharp videos from new viewpoints from a single motion-blurred image of a face. Our method[1] handles the complexity of face blur by implicitly learning the geometry and motion of faces through the joint training on three large datasets: FFHQ and 300VW, which are publicly available, and a new Bern Multi-View Face Dataset (BMFD) that we built. The first two datasets provide a large variety of faces and allow our model to generalize better. BMFD instead allows us to introduce multi-view constraints, which are crucial to synthesizing sharp videos from a new camera view. It consists of high frame rate synchronized videos from multiple views of several subjects displaying a wide range of facial expressions. We use the high frame rate videos to simulate realistic motion blur through averaging. Thanks to this dataset, we train a neural network to reconstruct a 3D video representation from a single image and the corresponding face gaze. We then provide a camera viewpoint relative to the estimated gaze and the blurry image as input to an encoder-decoder network to generate a video of sharp frames with a novel camera viewpoint. We demonstrate our approach on test subjects of our multi-view dataset and VIDTIMIT.

1   Introduction

Faces are a fundamental subject in image processing and in recognition due to their role in applications such as teleconferencing, video surveillance, biometrics, video analytics, entertainment, and smart shopping, just to name a few. In particular, in the case of teleconferencing, the interaction is found to be more engaging when the person on the screen looks towards the receiver [48]. However, to achieve this configuration it is necessary to look directly into the camera. Unfortunately, this does not allow to watch the person on the screen that one talks to. A solution to this issue is to design a system that can render the captured face from an arbitrary viewpoint. Then, it becomes possible to dynamically adapt the gaze

[1]Code and data available at https://gmeishvili.github.io/deblur_and_rotate_motion_blurred_faces/index.html.

Figure 1: **Blurry inputs and reconstructed sharp multi-view videos on our dataset (to play the videos open the paper pdf with Adobe Acrobat Reader).** We propose a model that, given the image of a blurry face, can render a corresponding sharp video from arbitrary viewpoints.
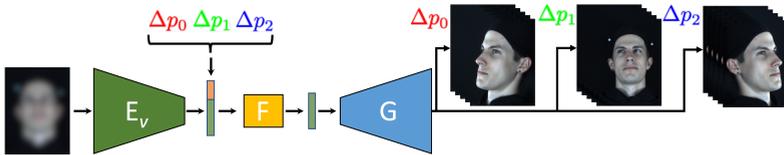

Figure 2: **Overview of our system during inference.** The encoder E_v encodes a blurry image into a sequence of latent codes that are then manipulated based on a relative viewpoint (*e.g.*, $\Delta p_0, \Delta p_1$, or Δp_2) via the fusion network F to produce encodings of images from a novel view. Finally, the generator G maps the novel view encodings to the image space.

of the face on the screen to ensure that it aims at the observer. Moreover, because of the low frame rate of web cameras, especially when used in low light, it becomes important to solve the above task in the presence of motion blur. Since a blurry image is the result of averaging several sharp frames [29], one could pose the problem of recovering not one, but a sequence of sharp frames from the single blurry input. This capability enables a smooth temporal rendering of the video. In addition, one might use this capability to deal with a limited connection bandwidth. Current software fits the available bandwidth by reducing the frame rate of the captured video. However, instead of selecting temporally distant frames, one could also transmit the average of several frames and then restore the original (high) frame rate at the destination terminal.

In this paper, we present a method that recovers a sharp video rendered from an arbitrary viewpoint from a single blurry image of a face (see Figure 1). Figure 2 shows our model during the inference stage. We design a neural network and a training scheme to remove motion blur from an image and produce a video of sharp frames with a general viewpoint. Our neural network is built in two steps: First, by training a generative model that outputs face images from zero-mean Gaussian noise, which we call the *latent space*, and then by training encoders to map images to the latent space. The primary motivation to use a generative model is that face rotations can be handled more easily in the latent space than in image space. This property was recently observed for generative adversarial networks [51]. One encoder is trained so that, when concatenated with the generator, it autoencodes face images. Then, rather than using sharp images as targets in a loss, we use their encodings, the latent vectors, as targets. As a second step, we obtain a blurry image by averaging several sharp frames. Then, we train a second encoder to map the blurry image to a sequence of latent vectors that match the target latent vectors corresponding to the original sharp frames. Finally, the change of the face viewpoint requires the availability of the latent vectors corresponding

to the same face instance, but rotated. To the best of our knowledge, there are no public face datasets with such data. Thus, we built a novel multiview face dataset. This dataset consists of videos captured at 112 fps of 52 individuals performing several expressions. Thanks to the high frame rate, we can simulate realistic blur through temporal averaging. Each performance is captured simultaneously from 8 different viewpoints so that it is possible to encode multiple views of the same temporal instance into target latent vectors and then train a fusion network to map the latent vector of one view and a relative viewpoint to the latent vector of another view of the same face instance. The relative viewpoint we provide as input should be the relative pose between the input and the output face poses. While we can use the viewpoint information from our calibrated camera rig during training, with new data, this information may be unknown. Hence, we also train a neural network to estimate the head pose. The network learns to map an image to Basel Face Model [31] (BFM) parameters, such that, when rendered (through a differential renderer), it matches the input image.

Contributions. We make the following contributions: *(i)* We introduce BMFD, a novel high frame rate multi-view face dataset that allows more accurate modeling of natural motion blur and the incorporation of 3D constraints; *(ii)* As a novel task enabled through this data, we propose a model that, given a blurry face image, can synthesize a sharp video from arbitrary views; *(iii)* We demonstrate this capability on our multiview dataset and VIDTIMIT [39].

2   Prior Work

3D Face Reconstruction. 3D morphable models (3DMM) [2] provide an interpretable generative model of faces in the form of a linear combination of base shapes. In the past decades many improvements were made using more data, better scanning devices or more detailed modelling [3, 4, 11, 25, 31, 32, 34, 36, 49, 50, 55]. 3D face reconstruction can be cast as regressing the parameters of such 3DMMs. The model parameters can be fit using multi-view images [33, 38, 40, 47, 52]. Since 3DMMs provide a strong shape prior, they also enable single-image 3D reconstruction [5, 23, 46]. These methods learn to estimate the model parameters by matching input images with differentiable rendering techniques [12, 22, 45, 60]. We also leverage a 3DMM to learn a controllable representation of faces. In our work, these representations are used to manipulate the latent space of a StyleGAN generator.

Face Deblurring. While we focus on modern learning-based approaches, there exist specialized classic approaches such as [30]. Several works designed specialized neural network architectures to target face deblurring. [8] performed face alignment to the input of the network. [9] introduced a two-stage architecture where the first stage restores low frequency and the second stage restores high-frequency content. [17] designed computationally efficient architecture that exploits a very large receptive field.

Some methods incorporate additional information in the form of semantic label maps [42, 56] or 3D priors from a 3DMM [57]. [26] disentangled image content and blur and exploit cycle-consistency to learn deblurring in the unsupervised, *i.e.*, unpaired setting. Face deblurring has also been combined with super-resolution by restoring high-resolution facial images from blurry low-resolution images [44, 53]. Our approach is different in that we invert a state-of-the-art generative model of face images.

Several works in the field focused on extracting a sharp video from a single motion blurred image [18, 35]. Jin et al. [19] introduced the task and solution of generating a sharp slow-motion video given a low frame rate blurry video.

Novel Face View Synthesis. Since our method allows the rendering of deblurred faces from
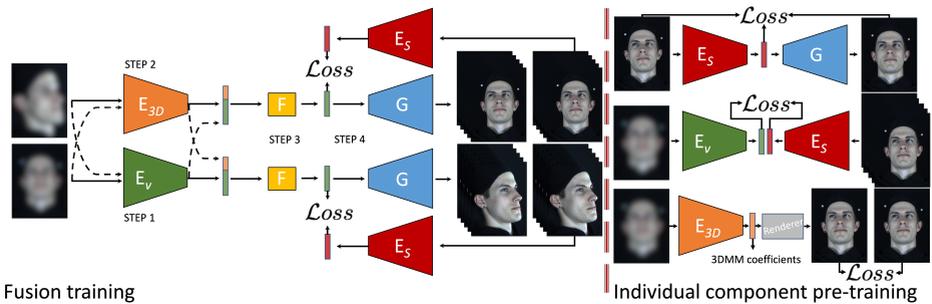
Figure 3: **Overview of the model architecture.** From top to bottom, on the right side of the figure, we show the individual pre-training stages of encoders: E_s, E_v, and E_{3D}. A sharp image generator G, is pre-trained using StyleGAN2. The training of the model F is shown on the right side of the figure. The encoder E_v encodes a blurry image into a sequence of latent codes corresponding to a sequence of sharp frames (step 1). Pose information is extracted via the viewpoint encoder E_{3D}, which is trained to regress the coefficients of a 3DMM (step 2). The predicted sharp latent codes are then manipulated based on the pose encodings via the fusion network F to produce latent codes of images from a novel view (step 3). Finally, generator G maps the novel view encodings to the image space (step 4).

novel views, we briefly discuss relevant work on novel face view synthesis. [54] use an encoder-decoder architecture. The encoder extracts view independent features, which are fed to the decoder along with sampled camera parameters. Realism and pose consistency are enforced via GANs. [15] use face landmarks to guide and condition the novel face view reconstruction. A special case of novel-view synthesis on faces is face frontalization [13, 27, 57]. [16] design a GAN architecture for face frontalization. Their generator consists of two pathways: A global pathway processes the whole image, and a local pathway processes local patches extracted at landmarks. Tackling the opposite problem, [58] train a GAN to generate silhouette images in order to reduce the pose bias in existing face datasets. To the best of our knowledge, we are the first to deblur and synthesize frames from a novel view simultaneously.

3   Model

Our goal is to design a model that can generate a sharp video of a face from a single motion-blurred image. Additionally, we want to synthesize novel views of these videos, *i.e.*, rotate the reconstructions. We design a modular architecture to achieve this goal (see Figure 3). We give an overview of the components here and provide more details in the following subsections. The bedrock of our approach is a generative model G of sharp face images. We describe how we can leverage the generative model G by learning an inverse mapping E_s from image-space to G's latent space in section 3.2. The sharp image encoder E_s then acts as a teacher for a blurry image encoder E_v. In section 3.3 we describe how to train E_v to predict latent codes of multiple sharp frames by using encodings of E_s as targets. To perform novel view synthesis, we require to capture the 3D viewpoint of the face. To this end, we learn a viewpoint extractor E_{3D} that maps a blurry image to coefficients of a 3DMM. We describe how to train E_{3D} using a differentiable renderer in section 3.4. The viewpoint from E_{3D} can then be used to manipulate the latent codes of a blurry image obtained through E_v. We do so by training a model F that, given relative viewpoint changes obtained through E_{3D} and
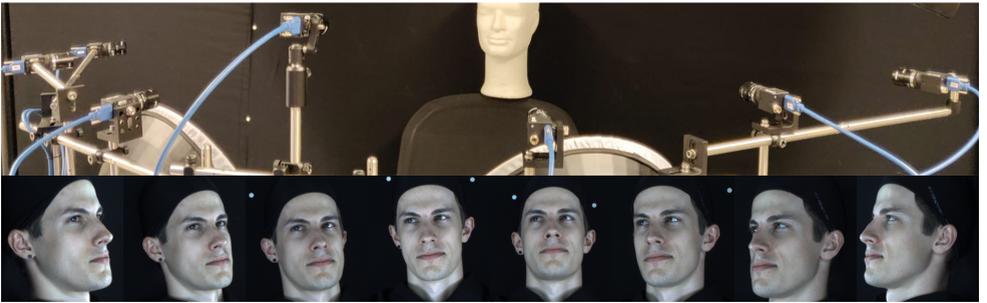
Figure 4: **Overview of our multi-view video capture setup.** In our lab setting, we arranged eight high-speed cameras in a circular grid. The cameras capture synchronized videos of participants performing a wide range of facial expressions from a wide variety of viewpoints. We show an example of 8 synchronized views of one of the 52 participants in BMFD. Background and clothing are black, allowing the easier extraction of skin regions.

latent codes from E_v, outputs updated latent codes corresponding to the desired change of viewpoint. This process is described in section 3.5.

Data. Our dataset consists of a set of sharp frames $\{y_i\}_{i=1}^{N}$. We synthesize blurry images by averaging $2m+1$ consecutive frames, $i.e.$, $x_i = \frac{1}{2m+1}\sum_{j=i-m}^{i+m} y_i$. As targets we define a sequence of 5 sharp frames $\mathbf{y}_i = [y_{i-m}, y_{i-m/2}, y_i, y_{i+m/2}, y_{i+m}]$. The training dataset then is given by $\mathcal{D} = \{(x_i^v, \mathbf{y}_i^v) \mid i = 1, \ldots, n; v = 1, \ldots, 8\}$, where the superscript v indicates the viewpoint (we omit v when it is not needed).

3.1 Bern Multi-View Face Dataset

Most prior face deblurring methods tackle the shift-invariant blur case, $i.e.$, blur that might arise from camera shake. Training data for such methods can be synthesized by convolving sharp face images with random blur kernels [17, 26, 42]. However, such models do not generalize well to blur caused by face motion since the resulting blurs are no longer spatially invariant. To tackle motion blur, Ren $et\ al.$ [37] generate training data by averaging consecutive frames of the 300-VW dataset [41]. This is a valid approximation of natural motion blur when the frame rate of the videos is sufficiently high. Since the 300-VW data has a relatively low frame rate of 25-30 fps, the resulting synthetic motion blurs are not always of high quality and can exhibit ghosting artifacts. Additionally, existing face datasets exhibit a pose bias, with most images showing faces in a frontal pose. Methods trained on such data can show poor generalization to non-frontal views.

To overcome these limitations, we introduce a dataset of high-speed, multi-view face videos. The faces of 52 participants were captured in a lab setting from 8 fixed viewpoints simultaneously. The cameras were arranged in a circular grid, ensuring that the faces are captured from all sides (see Figure 4). Videos are captured at 112 frames per second at a resolution of 1440×1080. The duration of the recordings ranges between 75 and 90 seconds.

3.2 Inverting a Generative Face Model

In order to generate novel views of a video sequence, we rely on a generative model of face images with a latent space where manipulations that change viewpoints are feasible. Consequently, we chose to train a SyleGAN2 [21] as the generator G of sharp face images.

SyleGAN2 provides state-of-the-art image quality and a smooth, disentangled latent space. To reconstruct or manipulate a given face image y_i, we require a corresponding latent code z_i, s.t. $G(z_i) = y_i$. To this end, we train a sharp image encoder E_s to invert the generator G, *i.e.*, we want that $G(E_s(y)) = y$. We adopt the inversion strategy of Meishvili *et al.* [28], where the encoder E_s is trained while the generator G is fine-tuned. The training objective is given by

$$\min_{E_s,G} \sum_{i=1}^{n} \ell_s \left(G\left(E_s(y_i) \right), y_i \right) + \lambda_g \left| G_{\text{init}} - G \right|_2^2 + \lambda_s \left| 1 - \left| E_s(y_i) \right| \right|, \qquad (1)$$

where ℓ_s represents the following combination of different reconstruction losses:

$$\ell_s (x,y) = \lambda_{id} \mathcal{L}_{id} (x,y) + \lambda_{per} \mathcal{L}_{per} (x,y) + \lambda_{edge} \mathcal{L}_{edge} (x,y) + |x - y|,$$

$$\mathcal{L}_{id} (x,y) = 1 - \frac{<\phi_{id}(x), \phi_{id}(y)>}{|\phi_{id}(x)| \cdot |\phi_{id}(y)|}, \quad \mathcal{L}_{per} (x,y) = \left| \phi_{per}(x) - \phi_{per}(y) \right|_2^2, \quad \mathcal{L}_{edge} (x,y) = \left| \mathrm{S}(x) - \mathrm{S}(y) \right|.$$

$\mathcal{L}_{id}$ is a term minimizing the cosine between embeddings of a pre-trained identity classification network ϕ_{id} of Cao et al. [7]. $\mathcal{L}_{per}$ is a perceptual loss on features of an ImageNet pre-trained VGG16 network ϕ_{per} [43]. $\mathcal{L}_{edge}$ is a Sobel edge matching term. We used a naive Bayes classifier with Gaussian Mixture Models trained on a skin image dataset from [1] to double the contribution of the skin pixels in all the losses.

$\lambda_g = 1$ controls how much G is allowed to deviate from the initial generator parameters G_{init} (before fine-tuning), and $\lambda_s = 1$ softly enforces that the predicted latent codes lie on the unit hypersphere. During training, we gradually relax λ_g until we reach the desired reconstruction quality. Similar to [28] we regress multiple latent codes per frame, each injected at different layers of the StyleGAN2. Thus $E_s(y_i) = z_i \in \mathbb{R}^{14 \times 512}$. Weights controlling the contribution of each term are set as follows: $\lambda_{id} = 0.5$, $\lambda_{per} = 10^{-6}$, $\lambda_{edge} = 0.2$, $\lambda_g = 1$.

3.3   Predicting Sharp Latent Codes from a Blurry Image

In this section we describe how to train a blurry image encoder E_v that maps a blurry image x_i to a sequence of 5 latent codes $\mathbf{z}_i = [z_{i-m}, z_{i-m/2}, z_i, z_{i+m/2}, z_{i+m}]$ corresponding to the target sharp frame sequence $\mathbf{y}_i$. We train the encoder E_v by using the the pre-trained sharp image encoder E_s as teacher. Let $\mathbf{z}_i = [E_s(y_{i-m}), \dots, E_s(y_{i+m})]$ denote the sequence of target codes obtained by encoding each target sharp image in the sequence $\mathbf{y}_i$ with E_s.

Jin *et al.* [18] point out ambiguities when regressing a sequence of sharp frames from a blurry image. Indeed, the order of the regressed frames can be ambiguous since the output sequence is often valid whether it is played forward or backward. We handle this forward/backward ambiguity by allowing for either solution in the training objective. Let the reversed target sequence be denoted with $\bar{\mathbf{z}}_i = [E_s(y_{i+m}), \dots, E_s(y_{i-m})]$. The training objective for E_v is then given by

$$\min_{E_v} \sum_{n=1}^{n} \min \left(|E_v(x_i) - \mathbf{z}_i|, |E_v(x_i) - \bar{\mathbf{z}}_i| \right), \qquad (2)$$

where we minimize either over the forward or backward target sequence, depending on which one better matches the prediction.

3.4   Regressing a 3D Face Model

To perform a novel view synthesis of the reconstructed sharp frame sequence, we need to know the 3D rotation of the face. Our approach is to learn to extract the 3D viewpoint of a

face by training an encoder E_{3D} to regress the coefficients of a 3DMM [31] along with camera parameters that define the rotation angles $R \in \mathbb{R}^3$, the translation $t \in \mathbb{R}^3$, and the illumination coefficients $\gamma \in \mathbb{R}^9$. The 3DMM coefficients can be grouped into components responsible for representing identity α, texture β, and facial expression δ. Given a blurry face image x_i^v from view v, we thus train a ResNet-50 [14] to regress the vector $c_i^v = (\alpha_i, \beta_i, \delta_i, \gamma_i^v, R_i^v, t_i^v) \in \mathbb{R}^{460}$ of 3D coefficients corresponding to the sharp middle frame y_i^v. The predicted 3D coefficients c_i^v are passed through a differentiable renderer ϕ [45] and the 3D encoder E_{3D} is trained by minimizing

$$\min_{E_{3D}} \sum_{i=1}^n \sum_{v=1}^{v_i} \ell_{im}\left(\phi\left(E_{3D}(x_i^v)\right), y_i^v\right) + \ell_{3D}\left(E_{3D}(x_i^v), y_i^v\right) + \lambda_c\left(|\alpha_i|^2 + |\beta_i|^2 + |\delta_i|^2\right), \quad (3)$$

where ℓ_{im} and ℓ_{3D} are a combination of different reconstruction losses (see supplementary for details), and $\lambda_c = 10^{-4}$ controls the amount of regularization applied to the 3DMM coefficients to prevent a degradation of face shape and texture. Note that the coefficients, α, β, γ, are shared across different views, promoting the accurate learning of facial expressions.

3.5   Learning to Rotate Faces in Latent Space

Given a blurry image x_i^v from viewpoint v and associated latent codes $z_i^v = E_v(x_i^v)$ as well as pose information $E_{3D}(x_i^v)$, we aim to manipulate z_i^v in latent space such that the reconstruction exhibits a desired change of viewpoint. We implement this by learning a fusion network F that takes as input a pair $(z_j^v, \Delta p)$ consisting of a single frame encoding z_j^v and a relative change in pose Δp. The output modified latent codes are then given by applying F to all frames in the sequence independently, i.e., the modified codes are given by $z_i^{v+\Delta p} = [F(z_{i-m}^v, \Delta p), \ldots, F(z_{i+m}^v, \Delta p)]$.

During training, we sample two blurry images x_i^u and x_i^v from two different viewpoints, but with the same timestamp. The change in viewpoint is then computed from $E_{3D}(x_i^u) - E_{3D}(x_i^v)$, which corresponds to $\Delta p_i^{uv} = (R_i^v - R_i^u)$, i.e., the difference in the estimated 3D rotation angles between the two views. We train the fusion model F to regress the latent codes z_i^v from the pair $(z_i^u, \Delta p_i^{uv})$ by optimizing the following objective

$$\min_F \sum_{i=1}^n \sum_{u \neq v} \min(|F(z_i^u, \Delta p_i^{uv}) - z_i^v|, |F(z_i^u, \Delta p_i^{uv}) - \bar{z}_i^v|), \quad (4)$$

where the `min` function again takes care of possible frame order ambiguities.

3.6   Implementation Details

We employed ResNet-50 [14] as a backbone architecture for E_s, E_v and E_{3d}. The average-pooled features are fed through fully-connected layers with 14×512 (single frame), $5 \times 14 \times 512$ (5 frames) and 460 neurons for E_s, E_v and E_{3d} respectively. The generator G is pretrained with all hyper-parameters set to their default values on 8 NVIDIA GTX 1080Ti GPUs (see Karras et al. [21] for details). All other networks were trained on 3 NVIDIA GeForce RTX 3090 GPUs. The Adam optimizer [24] with a fixed learning rate of 10^{-4} was used for the training of all the networks. We used batch sizes of 72, 96, 90, 84 samples for E_s, E_v, E_{3d} and F respectively. We trained our models E_s, E_v, E_{3d} and F for 1000K, 100K, 600K, and 500K iterations each. The ratio of samples within one batch stemming from FFHQ, 300VW and BMFD is 2:1:1. All the models are trained on an image resolution of 256×256. We used random jittering of hue, brightness, saturation, and contrast for data augmentation.

Frames	Views				
	1,8	2,7	3,6	4,5	All
Middle 3	2.93	2.78	2.89	2.82	2.85
Frames 2,4	3.31	3.13	3.29	3.27	3.25
Frames 1,5	3.94	3.84	3.99	4.01	3.95
All Frames	3.50	3.35	3.51	3.49	3.46

Table 1: **Same view landmark error.** We report the landmark error (in pixels) between the ground-truth and reconstructed frame sequences without rotation.

Fusion	Viewpoint Change		
	$\pm 30°$	$\pm 45°$	$\pm 60°$
FC3	51% (86%)	27% (62%)	14% (37%)
FC3R	56% (84%)	36% (66%)	22% (45%)

Table 2: **Identity agreement between frontal and rotated sequences.** We report the Top-1 (Top-5) label agreement of a pre-trained identity classifier between frontal and rotated views. Note that the classifier has a sensitivity of 61% (87%) on average over all viewpoints.

4   Experiments

Datasets. Besides our novel multi-view face dataset we also use 300VW [10], FFHQ [20] and VIDTIMIT [59] in our experiments. To synthesize motion-blurred images for training, we average *(i)* 65 consecutive frames from videos of 40 identities of our new dataset, and *(ii)* 9 consecutive frames from 65 identities of 300VW. To increase the number of identities for training and avoid overfitting, we also incorporate samples from FFHQ. Since FFHQ consists of still images, we simulate blurs by convolving images with randomly sampled 9×9 motion blur kernels. Because 300VW and FFHQ lack multiple views, we simulate them via horizontal mirroring of frames. We evaluate our method on the remaining identities of our new dataset and the VIDTIMIT dataset.

Pose-Regression Accuracy of E_v. We perform experiments to quantify the facial pose accuracy of the reconstructed frame sequence $G(E_v(x))$. To this end, we extract facial landmarks using the method of [6] from both the reconstructed and the ground-truth frame sequence on test subjects of our dataset. We report the MSE between them in Table 1 (again adjusting for the forward/backward ambiguity). We observe that the mean landmark error is slightly larger for peripheral frames (1, 2, 4, and 5) than the middle one (3). The mean landmark error is 3.46 pixels which amounts to 1.35% of the 256×256 image resolution.

Identity Preservation and Pose Accuracy under Novel View Synthesis. A key component of our method is the fusion model F, which performs the manipulation in the latent space that results in a change of the viewpoint. We thus perform ablation experiments for different architecture designs of F, where we measure how well they reconstruct the pose in novel views and how well they preserve the identity of the face. We consider two functional designs: *(i)* $FCxR$, where F is modelled via residual computation, *i.e.*, $F(z, \Delta p) = z + MLP_x([z, \Delta p])$, and *(ii)* FCx, where F simply consists of x fully-connected layers, *i.e.*, $F(z, \Delta p) = MLP_x([z, \Delta p])$ (x indicates the number of layers in the *MLP*). We want F only to affect the 3D orientation of the face in our method and preserve the face identity as much as possible. To quantify the consistency of face identities under novel view synthesis, we compute the agreement of a pre-trained identity classifier [7] between a restored frontal view and reconstructions under varying amounts of rotation. We report the resulting Top-1 and Top-5 label agreements on VIDTIMIT in Table 2. Because the identity classifier is not perfectly robust to face rotations, we also report the estimated identity agreement of the classifier (its sensitivity) on sharp ground-truth rotations. We observe that the identity labels of rotated sequences are relatively consistent with the classifier's sensitivity on ground truth rotations up to $\pm 30°$. The residual version $FCxR$ performs considerably better.

To quantify the accuracy of the predicted face pose under novel view synthesis, we mea-

blurry input   our result   sharp target   Zhou et al. [59]   sharp input
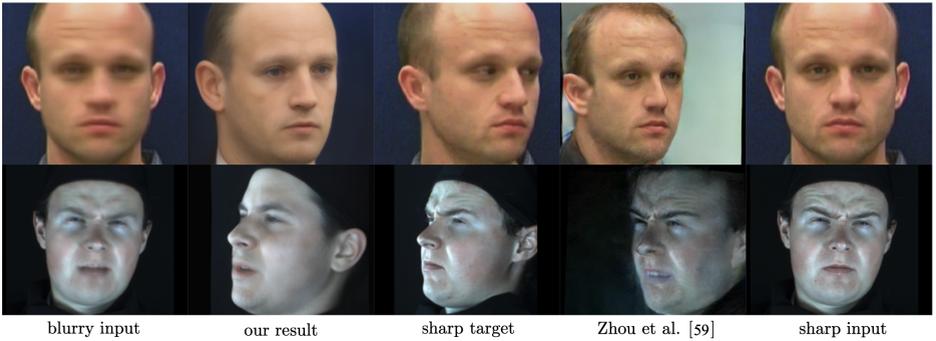
Figure 5: **Qualitative novel view comparison to Zhou et al. [59].** We compare on VID-TIMIT (top) and BMFD (bottom). Note that [59] predicts novel views from the sharp input image on the right, whereas we predict it from the blurry image on the left.

| Frames | Fusion | Views | | | | |
		1,8	2,7	3,6	4,5	All
Middle 3	FC3R	6.07	7.37	7.03	3.80	6.07
	FC3	6.67	7.51	7.02	3.99	6.30
Frames 2,4	FC3R	6.08	7.33	7.03	3.85	6.07
	FC3	6.61	7.49	6.99	4.02	6.28
Frames 1,5	FC3R	6.20	7.46	7.17	4.03	6.21
	FC3	6.63	7.63	7.14	4.20	6.40
All Frames	FC3R	6.12	7.39	7.09	3.91	6.13
	FC3	6.63	7.55	7.06	4.09	6.33

Table 3: **Face landmark accuracy for different fusion models.** In the table we report the landmark error of different frames in the reconstructed sequence (rows) and when faces are rotated to the different views in BMFD (columns). The blurry input image is taken from view 4 in all cases. An illustration of the frame and view layout is given on the right.

sure the face landmark error between the ground truth views and our reconstructions on test subjects of our multi-view dataset. Blurry frontal images (view 4) are fed through our model to reconstruct sharp frame sequences corresponding to the other seven views in our dataset. We report the mean landmark errors of different fusion models for all the views and predicted frames in Table 3. We observe that the average error across all views and frames varies between 6.13 and 6.33 pixels. Note that the reconstructions without rotations already show a mean landmark error of 3.46 pixels (see Table 1). Qualitative reconstructions of frontal and rotated frame sequences obtained with our method can be found in Figure 6.

Comparison to Prior Work. We compare to Zhou et al. [59] on novel face view synthesis quantitatively in Table 4 and qualitatively in Figure 5. Since [59] is trained on non-blurry face images, we feed it with sharp frontal views from VIDTIMIT and our test set. Our method was instead evaluated on blurry input images. Despite this disadvantage, our method yields a comparable accuracy. More results are shown in the supplemental material.

We evaluated the performance of our system using conventional metrics such as PSNR and SSIM. None of the existing prior deblurring work can generate novel views from a blurry input. Therefore, we use the combination of two methods for comparison purposes. We extract the sharp video sequence from a blurry input utilizing the method of Jin et al. [18] and subsequently rotate the resulting frames using the method of Zhou et al. [59]. The mean PSNR and SSIM between ground-truth and rotated sequences are reported in Table 5.
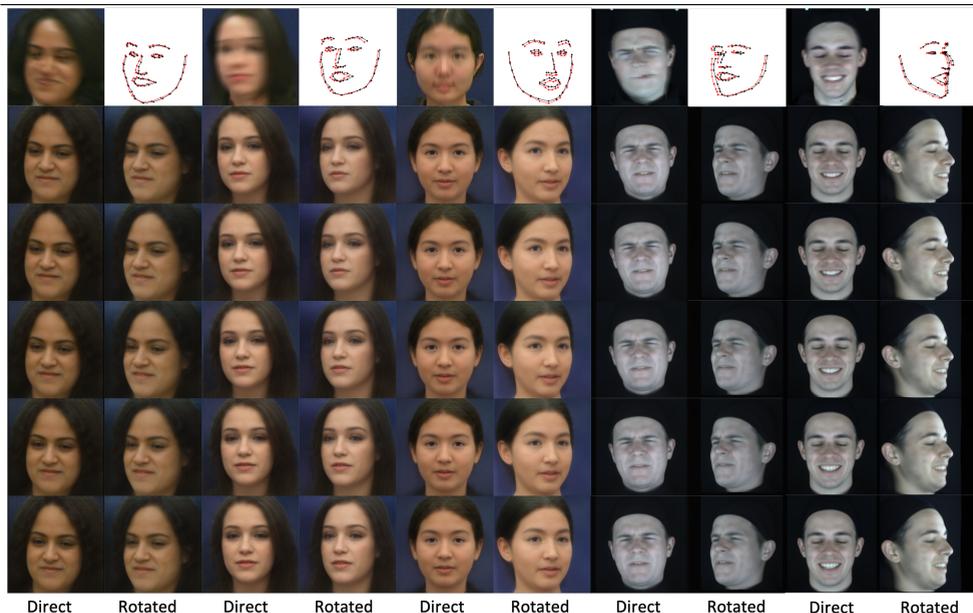
Figure 6: **Sample sharp video reconstructions from our model.** We show reconstructed frame sequences without viewpoint change (odd columns) and with random viewpoint changes (even columns). The first row shows the blurry input image followed by landmarks computed on the first and last frame in the reconstructed sequence. The first three examples are computed on VIDTIMIT and the last two on our test set.

Method	BMFD					VIDTIMIT
	1,8	2,7	3,6	4,5	All	
Zhou et al. [59]	7.12	6.42	5.40	5.61	6.14	3.12
Ours	6.07	7.37	7.03	3.80	6.07	3.96

Table 4: **Novel view pose error comparison.** We compare to the prior novel face view synthesis method by [59] in terms of face landmark accuracy on VIDTIMIT and BMFD.

Method	PSNR	SSIM
Jin et al. [18] + Zhou et al. [59]	16.07	0.38
Ours	19.45	0.60

Table 5: **Novel view PSNR and SSIM comparison.** We compare to the prior work in terms of PSNR and SSIM metrics on our dataset. First, the blurry input images from view 4 are fed to the method of Jin et al. [18], then, the resulting deblurred sequences are rotated using the method of Zhou et al. [59].

5   Conclusions

In this paper, we have presented a first method to reconstruct novel view videos from a single motion-blurred face image. Capabilities of the method were demonstrated on the VIDTIMIT dataset and a novel high frame rate, multi-view facial dataset, which we introduced. The multi-view dataset is crucial in enabling the training of our model. Moreover, our dataset is not limited to our proposed task: It can also be used to evaluate facial restoration methods for 3D reconstruction, single/video super-resolution, and temporal frame interpolation.

References

[1] Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1), 2002.

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999.

[3] Timo Bolkart and Stefanie Wuhrer. A robust multilinear model learning framework for 3d faces. In *CVPR*, 2016.

[4] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016.

[5] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models "in-the-wild". In *CVPR*, 2017.

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[8] Grigorios G. Chrysos and Stefanos Zafeiriou. Deep face deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[9] Grigorios G. Chrysos, Paolo Favaro, and Stefanos Zafeiriou. Motion deblurring of faces. *International Journal of Computer Vision*, 127(6), 2019.

[10] Grigoris G. Chrysos, Epameinondas Antonakos, Stefanos Zafeiriou, and Patrick Snape. Offline deformable face tracking in arbitrary videos. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.

[11] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models - past, present and future. *ACM Transactions on Graphics*, 39(5), August 2020.

[12] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, 2018.

[13] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[16] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[17] Meiguang Jin, Michael Hirsch, and Paolo Favaro. Learning face deblurring fast and wide. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[18] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. In *CVPR*, 2018.

[19] Meiguang Jin, Zhe Hu, and Paolo Favaro. Learning to extract flawless slow motion from blurry videos. In *CVPR*, 2019.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[22] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018.

[23] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *CVPR*, 2018.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Feng Liu, Luan Tran, and Xiaoming Liu. 3d face modeling from diverse raw scan data. In *ICCV*, 2019.

[26] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. Unsupervised domain-specific deblurring via disentangled representations. In *CVPR*, 2019.

[27] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*. IEEE, 2018.

[28] Givi Meishvili, Simon Jenni, and Paolo Favaro. Learning to have an ear for face super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.

[30] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring face images with exemplars. In *ECCV*, 2014.

[31] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, 2009.

[32] Weilong Peng, Zhiyong Feng, Chao Xu, and Yong Su. Parametric t-spline face morphable model for detailed fitting in shape subspace. In *CVPR*, 2017.

[33] Marcel Piotraschke and Volker Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *CVPR*, 2016.

[34] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William A. P. Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *CVPR*, 2019.

[35] Kuldeep Purohit, Anshul Shah, and A. N. Rajagopalan. Bringing alive blurred moments. In *CVPR*, 2019.

[36] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3d faces using convolutional mesh autoencoders. In *ECCV*, 2018.

[37] Wenqi Ren, Jiaolong Yang, Senyou Deng, David Wipf, Xiaochun Cao, and Xin Tong. Face video deblurring using 3d facial priors. In *ICCV*, 2019.

[38] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *CVPR*, 2016.

[39] Conrad Sanderson and Brian C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In Massimo Tistarelli and Mark S. Nixon, editors, *Advances in Biometrics*, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[40] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, 2019.

[41] Jie Shen, Stefanos Zafeiriou, Grigoris G. Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[42] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Yibing Song, Jiawei Zhang, Lijun Gong, Shengfeng He, Linchao Bao, Jinshan Pan, Qingxiong Yang, and Ming-Hsuan Yang. Joint face hallucination and deblurring via structure generation and detail enhancement. *International Journal of Computer Vision*, 2019.

[45] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv:1910.00287*, 2019.

[46] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018.

[47] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Perez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *CVPR*, 2019.

[48] Michael Tomasello, Brian Hare, Hagen Lehmann, and Josep Call. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *J Hum Evol*, 52(3), Mar 2007. ISSN 0047-2484 (Print); 0047-2484 (Linking).

[49] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018.

[50] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.

[51] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space, 2020.

[52] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, 2019.

[53] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[54] Xiaogang Xu, Ying-Cong Chen, and Jiaya Jia. View independent generative adversarial network for novel view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[55] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *CVPR*, 2020.

[56] R. Yasarla, F. Perazzi, and V. M. Patel. Deblurring face images using uncertainty guided multi-stream semantic networks. *IEEE Transactions on Image Processing*, 29, 2020.

[57] Zhihong Zhang, Xu Chen, Beizhan Wang, Guosheng Hu, Wangmeng Zuo, and Edwin R Hancock. Face frontalization using an appearance-flow-based convolutional neural network. *IEEE Transactions on Image Processing*, 28(5), 2018.

[58] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *Advances in neural information processing systems*, 2017.

[59] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[60] Wenbin Zhu, HsiangTao Wu, Zeyu Chen, Noranart Vesdapunt, and Baoyuan Wang. Reda:reinforced differentiable attribute for 3d face reconstruction. In *CVPR*, 2020.