

Efficient conditioned face animation using frontally-viewed embedding

Maxime Oquab^{*}, Daniel Haziza^{*}, Ludovic Schwartz^{†*}, Tao Xu^{*},
 Katayoun Zand^{*}, Rui Wang^{*}, Peirong Liu^{◇*}, Camille Couprie^{*}
^{*} Meta, [†] Universidad Pompeu Fabra, [◇] UNC-Chapel Hill

{qas, dhaziza, xutao, couprie}@fb.com

Abstract

As the quality of few shot facial animation from landmarks increases, new applications become possible, such as ultra low bandwidth video chat compression with a high degree of realism. However, there are some important challenges to tackle in order to improve the experience in real world conditions. In particular, the current approaches fail to represent profile views without distortions, while running in a low compute regime. We focus on this key problem by introducing a multi-frames embedding dubbed Frontalizer to improve profile views rendering. In addition to this core improvement, we explore the learning of a latent code conditioning generations along with landmarks to better convey facial expressions. Our dense models achieves 22% of improvement in perceptual quality and 73% reduction of landmark error over the first order model baseline on a subset of DFDC videos containing head movements. Declined with mobile architectures, our models outperform the previous state-of-the-art (improving perceptual quality by more than 16% and reducing landmark error by more than 47% on two datasets) while running on real time on iPhone 8 with very low bandwidth requirements.

1. Introduction

Despite the progress of recent video codecs, they fail to provide meaningful videos in ultra low bandwidth situations. To bridge this gap, recent generative approaches [8, 11, 17] suggested to compress and send facial landmarks with a reference image, and animate it using variants of the First Order Model (FOM) work [15].

To be used for video chat, a generative low-bandwidth approach needs to satisfy several constraints. Some are related to inter-person communication, *e.g.* identity preservation without uncanny valley effects, and proper fidelity of facial motion. Some are related to the technical context, *i.e.* the algorithm must work in real-time, preferably on mobile

^{*}Work achieved during internship at Meta.

devices, and use only little bandwidth.

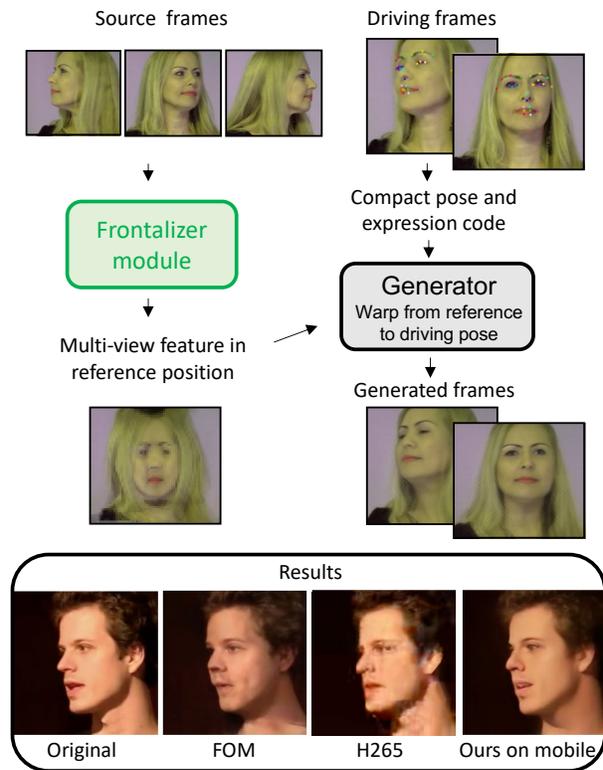


Figure 1. We introduce the Frontalizer module, that builds a unified embedding from a set of source frames, allowing our generator based on the First Order Model [15] (FOM) to use appearance information from multiple frames. While in the algorithm, this happens in **feature space**, for this figure we apply the same transforms to the source images for illustrating the multi-view feature. Our resulting face animation approach is running on real time on iPhone8 at less than 7 kbps (kbits per second) using three source frames, improving over FOM and the HEVC codec. The transmission of two additional frames is amortized in a few seconds.

Leading face animation approaches are based on the FOM work by deforming a source image to match the position of a driving image. However, this approach suffers



Figure 2. Reconstruction of a side view by a **baseline FOM** [15] **model** using one frontal source frame. From left to right: source image with keypoints; driving image with keypoints; reconstruction with a dense model; reconstruction with a mobile model. The side-view appearance information is absent, and the results present noticeable distortions. The effect is stronger on mobile where the capacity is low.

from one fundamental issue: the hidden areas, that were occluded in the source image, are implicitly inpainted by an image generator model at inference time. This can lead to small perceptible differences in the result and an uncanny valley effect, as it can be seen in the FOM result of Figure 2. When approaching the problem with mobile models, the reconstruction quality for these areas decreases even more as the capacity of the model is reduced and the result becomes unacceptable for a video call. In order to tackle this problem in a small model regime, it is critical to use multiple views of a face. Combining these views efficiently is a non-trivial problem. One possible approach would consist of dynamically switching between several source frames and picking the one closest to the driving position. However, for the result to appear natural and temporally consistent, it is important to avoid abrupt changes in the appearance as can happen when switching from one source frame to another. Therefore, our first contribution is the introduction of a Frontalizer module, that combines a set of source frames into a single reference embedding map, containing the information from all views of the source set in feature space, see Figure 1.

Then, in order to refine the result, we introduce a lightweight latent code that is transmitted along with the driving position. This code is given to the generator network and optimized to improve the reconstruction. We show that this can refine difficult areas such as the inside of the mouth.

Finally, as the goal of this work is to enable a generative-based video-chat, we propose a mobile architecture and implementation that runs at 30 FPS on a 2017 iPhone 8, leveraging GPU hardware acceleration, while constantly outperforming the state-of-the-art on two datasets.

2. Related work

Facial animation. The First Order Model of Siarohin *et al.* [15] provides a strong baseline for face animation models. Unlike prior work [18, 21], it only requires one source frame to accurately reconstruct the frontal view of talking heads. However, solely using one face image makes the re-

construction of profile views extremely challenging, even more so in a low computation regime.

Different attempts appeared to cope with this problem. In [21], a finetuning step using a minimum of height frames compensates the identity loss in the approach, that would be tricky to implement on device. In the Bilayer approach [20], the source image is first transformed to extract a high frequency feature embedding, then used to compute the feature deformation field. A drawback of high/low frequency differentiation is the presence of hair artefacts. In [17], it is the use of 3D landmarks that aims to provide accurate profile views. However the rotation angle remains limited and the very high dimension of necessary feature space for this approach prevents immediate application to real time on device inference. Closer to our solution, the work of Wiles *et al.* [18] embeds multiple frames using a single encoder, but then deforms images directly and not their features, leading to limited performance. Audio based approaches such as [24] animate faces from audio features, a source image and driving poses. [24] demonstrates convincing profile views rendering, however the code necessary to condition the generator would be too large (several thousands of floats) to make this approach work in a low bandwidth situation. Finally, the very recent work of [9] leverages multiple frames by fusing warped features from N multiple views in the driving domain. At inference time, it requires N extra calls to a warping network, making this approach unfit to low compute applications.

Expression conditioned models. Only conditioning generation on facial landmarks is not enough information to model some facial attributes that can be either occluded (such as the tongue and teeth) or ignored (e.g. wrinkles) by classical perceptual losses. In the literature, we only found approaches conditioning generations on facial expressions [12] or facial landmarks (see the previous section), but not both jointly.

Applications to low bandwidth video chat. Last year simultaneously appeared three generative approaches with applications to low bandwidth video chat: [8, 11, 17]. The transfer of Wang *et al.*'s approach discussed above [17] to low compute regime has not been studied so far, but the dimensionality of the latent space employed in their work is height times higher than the one of FOM, and 32 times higher than the one of [11]. Additionally, the use of 3D convolutions adds another level of complexity that would cause issues in low compute scenarios. It is therefore unlikely to be a system applicable on device. A server based solution is evoked in [1], where the reconstructed video of a sender with poor network connection could be sent to a receiver to only enable one way communication: The low bandwidth context would prevent the sender to receive their

interlocutor’s video. The work of Konuko *et al.* [8] is fully orthogonal to ours, as it introduces an algorithm to select intra frames to best compress videos using the original FOM algorithm. Finally, [11] presents a variant of FOM that runs on mobile, where the decoder contains SPADE layers to refine some areas of the face. Like FOM, it does not handle profile views generation well. Moreover, the addition of keypoints required by SPADE augments the bandwidth.

3. Model

We describe the model that we propose for low-bandwidth video-chat on mobile phones. In particular we explain which design choices were made to comply with constraints related to the task of low-bandwidth video-chat.

3.1. Challenges of face animation

3.1.1 Identity preservation

In order to provide a good video-chat experience, it is necessary that the identity of the caller is preserved. While it is difficult to computationally measure identity preservation in image generation, we observe that the First Order Model, based on warping feature maps, offers the best quality on that aspect as it preserves many details of the original face; we build on the FOM approach for the rest of this work.

First Order Model. We provide a brief description of FOM below and refer the reader to the original paper [15] for more details. For a source frame x_s and a driving frame x_d , an unsupervised keypoint detector \mathbf{Kp} computes a set of keypoints and their local Jacobian matrices, defining a source position p_s (and similarly, a driving position p_d):

$$p_s = \mathbf{Kp}(x_s) \quad ; \quad p_d = \mathbf{Kp}(x_d).$$

An encoder network \mathbf{Enc} computes the feature map embedding for the source frame in position p_s :

$$e^{p_s} = \mathbf{Enc}(x_s).$$

Next, a dense motion network \mathbf{Dm} constructs a flow transformation map $t^{p_s \rightarrow p_d}$ and an occlusion map $o^{p_s \rightarrow p_d}$ in order to build the feature map embedding in position p_d :

$$(t^{p_s \rightarrow p_d}, o^{p_s \rightarrow p_d}) = \mathbf{Dm}(x_s, p_s, p_d).$$

Then, the flow is applied onto the feature map using the warping “grid-sample” operator \mathbf{Gs} , before applying the occlusion map by element-wise multiplication \odot :

$$\tilde{e}^{p_d} = \mathbf{Gs}(e^{p_s}, t^{p_s \rightarrow p_d}) \odot o^{p_s \rightarrow p_d}.$$

Finally, a decoder network \mathbf{Dec} provides the result image:

$$\tilde{x}_d = \mathbf{Dec}(\tilde{e}^{p_d}).$$

3.1.2 Inpainting missing areas

One fundamental problem in warping-based approaches is their reliance on inpainting areas that were not visible in the original image. While it seems natural for a video-chat scenario to use a front-facing source frame, this means that the sides of the face will not be visible. In practice this means that when a head rotation happens, the side-view result image will be partially generated by the model, building on what was seen by the model in the training dataset. At inference time, on a new unseen user, the result will not match, leading to an *uncanny valley effect* (see Figure 2). Moreover, as this implicit inpainting relies on the *capacity* of the model to learn corresponding patterns in the training dataset, this problem is even more pronounced in a small model regime as is the case for mobile inference purposes, as detailed further in Section 3.3. We approach this inpainting problem with our Frontalizer module described in Section 3.2.1.

3.1.3 Expression preservation

One other problem we observe is the one of facial expressions preservation during the video-chat: during communication, facial expressions appear as a combination of motion of facial areas (e.g. eyebrows raising or frowning) and short displays of the corresponding facial wrinkles (e.g. horizontal forehead lines or vertical line between eyebrows). However, flow-based methods such as the FOM or its proposed variant in [17] (see Figure 6) tend to have trouble reproducing the latter. We approach this problem with our expression conditioning setup described in Section 3.2.2.

3.2. Proposed approach

An overview of the full pipeline for our approach appears in Figure 3. Our work introduces two main components: the Frontalizer module contains a first learnable sub-network, similar to the dense-motion network: its role is to produce a flow map from a source position to a reference position p^r . The reference position p^r corresponds to a set of K keypoint positions; we detail this component in Section 3.2.1.

The second component is an expression module, where a small CNN extracts from the driving frame a compact expression code to refine the result. This expression code is used by the decoder network as conditioning, in the form of one-hot maps. We detail this component in Section 3.2.2.

3.2.1 Frontalized embedding warp and fusion

The frontalizer module replaces the encoder module of the First Order Model. Given N source frames x_1, \dots, x_N in positions p_1, \dots, p_N , its goal is to compute and warp their feature maps into a reference position p_r in feature space, then

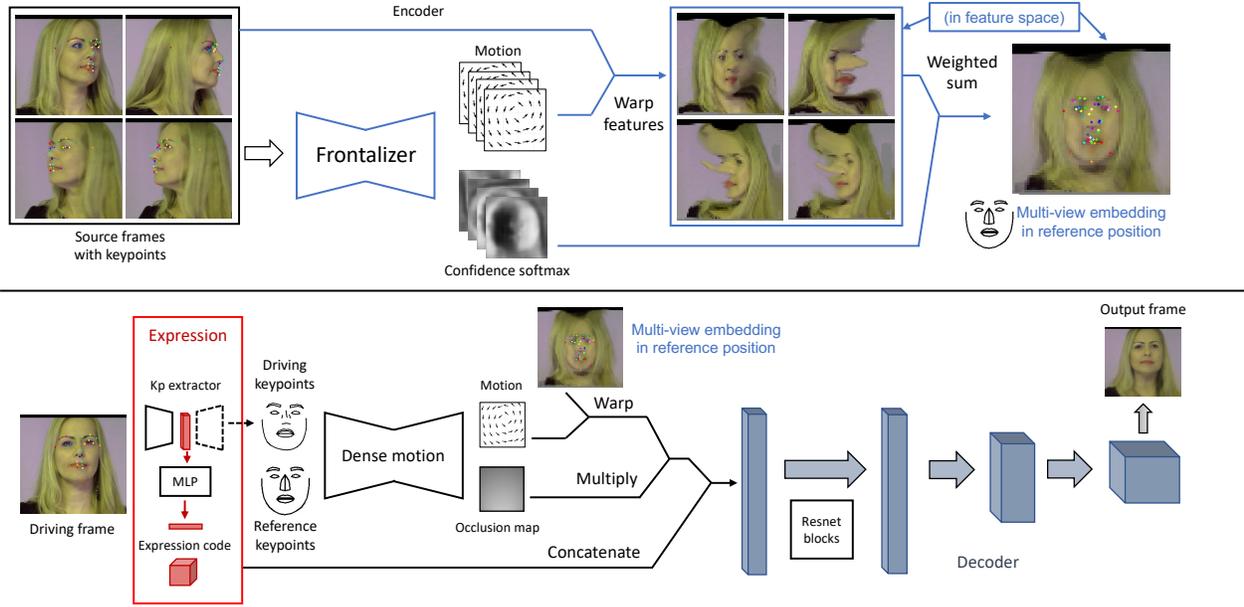


Figure 3. Overview of our approach. Top part: the Frontalizer first builds a face appearance embedding from a few source images, in a reference position. For the purpose of this figure, we illustrate the features by applying the transforms on the RGB frames to gain visual intuition. In the algorithm, these transforms are applied on the encoder feature maps. Bottom part: We use the multi-view embedding in the reference position to decode the result in a new driving position, hence reconstructing the driving frame only using the driving keypoints and an expression code. Finally, our decoder reconstructs the frame.

combine them together depending on their individual confidence maps c_1, \dots, c_N . We illustrate this component in the top part of Fig. 3, and detail its steps below:

First, we apply an encoder Enc to obtain the feature map for each image x_i at the original position p_i :

$$e^{p_i} = \text{Enc}(x_i).$$

Then, we compute a flow map $t_i^{p_i \rightarrow p_r}$ and a confidence map $c_i^{p_i \rightarrow p_r}$ using a dense motion network Dm re-purposed for frontalization, denoted Fr :

$$(t^{p_i \rightarrow p_r}, c_i^{p_i \rightarrow p_r}) = \text{Fr}(x_i, p_i, p_r).$$

We apply this flow to the feature maps to obtain the warped result in position r , using the grid-sample operator Gs :

$$\tilde{e}_i^{p_r} = \text{Gs}(e^{p_i}, t^{p_i \rightarrow p_r}).$$

Next, we compute weighting coefficient maps w_i for each flow map by applying a softmax operation to the N confidence values at each coordinate position (u, v) :

$$w_i(u, v) = \text{softmax}_i(c_i^{p_i \rightarrow p_r}(u, v)) \quad \forall i \in \{1, \dots, N\}.$$

Finally, we combine the different maps with an element-wise multiplication together to obtain the fused embedding:

$$\tilde{e}^{p_r} = \sum_{i=1}^N w_i \tilde{e}_i^{p_r}.$$

This fused embedding \tilde{e}^{p_r} and its reference position p_r are ready to be used by the generator (dense-motion and decoder networks) of a FOM setup. As it contains information from multiple views, the reliance on the inpainting properties of the decoder is now greatly reduced. We note that the learnable parts of the Frontalizer module consist of a dense motion network, that specializes in producing flow maps towards a reference position p_r (also learned), as well as corresponding confidence maps, optimizing the merging of the different sources.

Losses. We use the same loss terms as in the FOM work [15], in particular, the multi-scale perceptual loss and adversarial training.

3.2.2 Expression conditioning

To convey important facial expressions such as wrinkles and so on, it is essential to introduce a code to transmit in addition to the landmarks.

We introduce a vector $e \in \mathbb{R}^M$, of size $M = 16$ by default, aiming to store expression information. We feed this code to our model in input to the decoder. To concatenate its values to the input feature map of the decoder, we fill E_k maps with the spatial repetition of each value e_k , $k \in \{1, \dots, M\}$. To ensure that the given code serves the recovery of desired

facial expressions, we introduce a facial action unit loss using predictions from a network pre-trained to predict landmarks with its branch f_L , and to classify facial action units with branch f_E [3, 16, 19]. The loss enforcing expressions is defined as follows:

$$\mathcal{L}_E(x_d, \tilde{x}_d) = \gamma_E (\|f_E(\tilde{x}_d) - f_E(x_d)\|_1 + \|f_L(\tilde{x}_d) - f_L(x_d)\|_1),$$

where $\gamma_E \in \mathbb{R}^+$ and \tilde{x}_d is the output generation by our model from a driving image x_d . To predict the expression code, we take advantage of the landmark extractor Kp encoder f_m , and define a two-layer Perceptron g on top of it:

$$e = g(f_m(x_d)).$$

As the learning of e is only supervised by the loss on the final image (expression \mathcal{L}_E in addition to other terms, e.g. perceptual, adversarial), we denote this expression conditioning as free. We also experiment two variants: taking $e = f_E(x_d)$, referred as Oracle, and alternatively further supervising the learning of e using oracle prediction. The free strategy remains the best as shown in our ablation study.

3.3. Setup details

Architecture. Our models contain several sub-networks that are common with the original FOM work [15]: an encoder, a keypoint detector, a dense-motion network and a decoder network. For these networks, we refer the reader to the original paper for their architecture definitions. Architecture details of the Frontalizer and expression module, that are specific to this work, are given in the Appendix.

Landmark sets. We experiment with two variants of keypoint sets: supervised landmarks setups (Sup) use a set of 33 facial landmarks using a facial landmark tracker. Unsupervised setups (Unsup) use 10 unsupervised keypoints obtained with the trained keypoint detector, while Sup+Unsup use both. Frontalizer models trained only with the Unsup set proved difficult to train. Our hypothesis is that the initialization of the reference keypoints positions (see next section) is important, and it is unclear how to initialize the unsupervised set.

Training. We use the DFDC dataset [5] for training, as it provides videos with a wide range of head rotations, which is not the case for the VoxCeleb dataset (see Figure 4). For training Frontalizer setups, we sample the dataset examples in the following way: given a video of a person talking, we run an off-the-shelf face detector [23], and crop a square box in the video that contains the detected head for over 90 consecutive frames to obtain a *video tube* with a fixed background. We then run a head angle detector on the frames of the tubes, returning their yaw (vertical axis rotation angle). We sample three frames A, B, C where

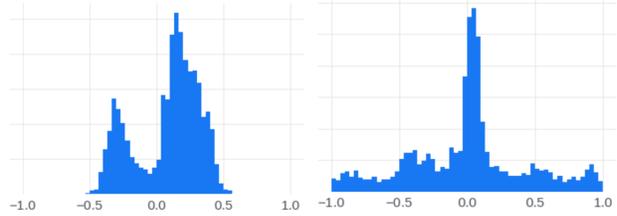


Figure 4. Yaw distributions (radians) for the test videos in VoxCeleb (left) and DFDC (right). Average (across videos) standard deviation of the yaw angle: VoxCeleb: 0.08; DFDC: 0.29. VoxCeleb videos contain only little head rotations compared to DFDC.

$\text{yaw}_A > \text{yaw}_B > \text{yaw}_C$ and $(\text{yaw}_A - \text{yaw}_C) > 0.3$ rad. We use A and C as source frames, and B as driving frame.

While this sampling strategy works well in presence of large head rotations, it neglects identity preservation in videos with less motion. To cope with that, we propose the following improved sampling using four frames A, B_1, B_2, C where

$$\text{yaw}_A > \text{yaw}_{B_1} > \text{yaw}_C,$$

$$\text{yaw}_A - \text{yaw}_C > 0.3 \text{ rad},$$

$$|\text{yaw}_A - \text{yaw}_{B_2}| < 0.1 \text{ rad}.$$

We use A and C as source frames, and B_1, B_2 as driving frames. This sampling allows the model to simultaneously learn from a large rotation regime (B_1 far from A and C) and a small rotation regime (B_2 close to A).

The Frontalizer reference keypoints are initialized, for the part corresponding to supervised facial landmarks, to the average position in the dataset, which is a front-facing layout. The reference unsupervised keypoints are initialized randomly close to the center. These reference keypoint positions are learned end-to-end during training.

Losses parameters. We use the Adam [7] optimizer ($\beta_1 = 0.5, \beta_2 = 0.99$), with all learning rates set to $2 \cdot 10^{-4}$; we divide the learning rate by 10 after 60k and 90k iterations, and use a batch size of 32 in all our ablation study experiments. We stopped training after 100k iterations, and fine-tune the FOM model with an adversarial loss [6] for 50k more iterations as in [15].

Mobile models. In order to run on mobile in real-time, the networks compute requirements and capacity must be greatly reduced. As mentioned in the introduction, decreasing the size of the model diminishes the inpainting capabilities of the decoder network. When the capacity is reduced to fit on mobile, this implicit inpainting becomes very difficult in practice - in particular, when there are head rotations - and the Frontalizer approach becomes extremely useful.

To enable mobile inference in real-time, we build all network architectures using efficient MobileNetV2 [14] inverted residual building blocks and reduce the size of the central feature from 64×64 to 32×32 . We provide architecture details in the Appendix.

4. Experiments

Datasets. We evaluate our models on two datasets to ease comparisons with previous approaches: VoxCeleb [10] and DFDC [5]. As shown in Fig. 4, VoxCeleb contains only little rotations on yaw angle; therefore we extract from DFDC a validation set of 157 videos, comprising 90 to 350 frames each, filtered to contain a maximum of head movements ($\max_{\text{yaw}} - \min_{\text{yaw}} > 1.5$ rad), and call it “DFDC-rotations”. We also use the DFDC-50 diverse set shared by [11].

Metrics. We provide different measures to assess image generation: The LPIPS [22] is a perceptual metric rating the overall quality of generations by comparing their deep features using a pre-trained network on image classification. Facial landmarks displacements are assessed by the NME as in [2]. To evaluate identity preservation, we rely on the CSIM metric, computing differences between Arcface features [4]. Finally, to assess facial expressions preservation that can not be computed from landmarks only, we compute an ℓ_1 norm of the difference of prediction of sixteen facial action units using a pre-trained network [16] to perform such classification. We denote this new metric as Expr.

Frame selection setup At inference time, we select one source frame having the smallest yaw angle (frontal view) over the video, and ensure that eyes are open. For the Frontalizer models, we select two additional frames where the yaw angles are the extrema reached in the video. Examples of selected source frames are shown in Figure 5.

We demonstrate the quality improvements brought by our models on dense architectures in Section 4.1, before presenting results on mobile architecture in Section 4.2. Finally, we quantify bandwidth requirements and compare to the previous on device state-of-the-art [11] in Section 4.3.

4.1. Dense models

We first provide qualitative results in Figure 5 that compares generations of dense models: The First Order Model (FOM adv), our result using a Frontalizer module with only supervised landmarks, using additional self-supervised landmarks, and using additional free expression conditioning. The progresses brought by the frontalization module, with respect to the FOM baseline are very significant. Indeed, the FOM baseline, having only access to a source image with frontal pose, can not recover occluded face parts accurately. By design, our Frontalizer variants, thanks to

their three source based embedding, infer realistic results. Among the Frontalizer variants, differences are less visible, however we note a better head shape preservation using the Sup+Unsup landmark setting, and a better background preservation using the free expression code. More examples are provided in Appendix.

Table 1 provides quality metrics for the FOM model and our Frontalizer variants on DFDC-rotations and Voxceleb. The frontalizer greatly outperforms FOM on DFDC-rotations. As mentioned previously, the vast majority of VoxCeleb videos do not contain head rotations, and when they do, these are limited to low yaw angles. In this context, it is not surprising that FOM reaches comparable performance. Using our novel sampling strategy, our approach manages to perform on par with FOM on VoxCeleb, even slightly improving it.

We highlight that on mobile (cf. Table 3), our models greatly outperform FOM on all datasets. The average NME of all approaches is in the range 0.247 - 0.256, indicating the degree of easiness of VoxCeleb for face animation. On DFDC-rotations, the diversity of head poses to render leads to a very high NME of 1.11 for FOM, and is much more reasonable for Frontalizer models, reaching 0.467 for the best model conditioned on expressions. The CSIM and LPIPS are further improved using additional Unsup landmarks, and the Expr metric using free expression conditioning.

Ablation study for the expression module. We present in Table 2 different variants of the expression conditioning we tried on top of the FOM baseline. First, we check that the addition of the Facial action unit loss alone does not help. Second, we compare expression conditioning by feeding “oracle” expression code in either the Dense motion network or in the Decoder. Adding these conditioning leads to slightly worse LPIPS but better expression metrics, with a larger improvement in expression brought by decoder conditioning. The boost brought by the light supervised code conditioning is small. Finally, the free light conditioning strategy is cheap to allow for mobile inference, does not increase the LPIPS significantly, and brings an improvement to the expression metric, justifying selecting this conditioning strategy for our experiments.

Qualitatively, as shown in Figure 6, our free conditioning helps to recover a good part of the wrinkles of the driving frames, while reducing the presence of hair artifacts in the background. It is possible to navigate in the free latent space and observe the effect of forcing different values for conditioning code. In Figure 7, setting some values of the code to zero or one leads to various mouth interior appearances.

Quantitative evaluation of dense models. To ensure that the findings given by automated metrics correspond to human perception, we conducted a human annotation of the



Figure 5. Our dense Frontalizer variants reconstruct faces accurately where the FOM adv baseline struggles with occluded parts given its only source input x_1 . The Frontalizer models also use the rightmost x_2 and leftmost x_3 frames.

	DFDC-rotations				VoxCeleb			
	LPIPS ↓	NME ↓	CSIM ↑	Expr ↑	LPIPS ↓	NME ↓	CSIM ↑	Expr ↑
FOM adv	0.136 \pm 0.004	1.108 \pm 0.176	0.784 \pm 0.008	0.854 \pm 0.008	0.118 \pm 0.003	0.247 \pm 0.007	0.826 \pm 0.008	0.915 \pm 0.003
Ours (Frontalizer sup)	0.132 \pm 0.002	0.496 \pm 0.034	0.793 \pm 0.005	0.879 \pm 0.005	0.171 \pm 0.004	0.269 \pm 0.004	0.789 \pm 0.005	0.897 \pm 0.004
Ours (Frontalizer sup+unsup)	0.113 \pm 0.002	0.469 \pm 0.029	0.799 \pm 0.005	0.881 \pm 0.005	0.151 \pm 0.003	0.255 \pm 0.004	0.791 \pm 0.005	0.896 \pm 0.004
Ours (Frontalizer sup+unsup + free expr)	0.113 \pm 0.002	0.467 \pm 0.030	0.789 \pm 0.005	0.897 \pm 0.004	0.140 \pm 0.003	0.256 \pm 0.003	0.781 \pm 0.005	0.910 \pm 0.004
Ours (Frontalizer sup+unsup) new sampling	0.088 \pm 0.002	0.425 \pm 0.027	0.850 \pm 0.004	0.902 \pm 0.004	0.116 \pm 0.002	0.237 \pm 0.002	0.837 \pm 0.002	0.915 \pm 0.002

Table 1. Quality metrics of dense models. All models were trained on DFDC using adversarial training. The baseline FOM setup compares favorably on the VoxCeleb dataset, which present only little rotations, as shown in Fig. 4. However our models compare favorably when large head angle rotations occur, as in the DFDC-rotations set.



Figure 6. Ablation study of expression conditioning. Free decoder conditioning keeps the wrinkles while limiting head distortions.



Figure 7. Variation in the free expression latent space (a): using original driving code e . (b) Setting e_7 to 1; (c) Setting e_6 to 0.

quality of reconstructed videos using the four models of Table 1, showing videos from DFDC-rotations and DFDC-50 to 15 annotators. 12% of FOM videos were preferred, 20% of Frontalizer Sup, 46% of Frontalizer Sup+Unsup, and 22% of Frontalizer S+U with expression free condition-

	LPIPS ↓	CSIM ↑	Expr ↑
FOM	0.115	0.841	0.934
FOM + FAU loss	0.119	0.822	0.926
FOM + DM expr cond.	0.117	0.833	0.951
FOM + Dec expr cond.	0.118	0.834	0.964
FOM + light Dec expr cond.	0.118	0.834	0.939
FOM + light Dec free cond.	0.116	0.834	0.949

Table 2. Ablation study for expression conditioning on DFDC-50. The free Decoder conditioning strategy is the best compromise.

ing.

4.2. Mobile models

In this section, we provide results obtained with mobile models. We replace the dense networks with their mobile counterparts, using MobileNetV2 building blocks, and the same network architectures for our approach and for the FOM-Mobile baseline, based on [15] and using ten unsupervised keypoints with their Jacobians. We also reduce the size of the bottleneck feature from 64×64 to 32×32 to save on computation and ensure a 30 FPS frame rate on iPhone 8 for all models. We report our results in Table 3. We observe that for both VoxCeleb and DFDC, in the mobile computation regime, our models using the Frontalizer compare favorably to the baseline FOM-Mobile, for image qual-

	DFDC-rotations				VoxCeleb			
	LPIPS ↓	NME ↓	CSIM ↑	Expr ↑	LPIPS ↓	NME ↓	CSIM ↑	Expr ↑
FOM-Mobile	0.187 ±0.004	1.122 ±0.105	0.722 ±0.005	0.790 ±0.006	0.197 ±0.002	0.519 ±0.007	0.717 ±0.004	0.837 ±0.003
Ours (sup)	0.172 ±0.003	0.632 ±0.062	0.710 ±0.006	0.866 ±0.005	0.184 ±0.002	0.288 ±0.003	0.736 ±0.003	0.890 ±0.002
Ours (sup+unsup)	0.148 ±0.003	0.585 ±0.044	0.723 ±0.006	0.865 ±0.005	0.165 ±0.002	0.275 ±0.003	0.747 ±0.003	0.893 ±0.002
Ours (sup+unsup)*	0.142 ±0.003	0.493 ±0.035	0.793 ±0.005	0.851 ±0.005	0.175 ±0.002	0.250 ±0.003	0.799 ±0.003	0.883 ±0.002

Table 3. Quality metrics of mobile models. *: model with new sampling and SPADE based decoder as in [11]. The new sampling strategy combined to SPADE layers improves results, particularly in terms of identity preservation.

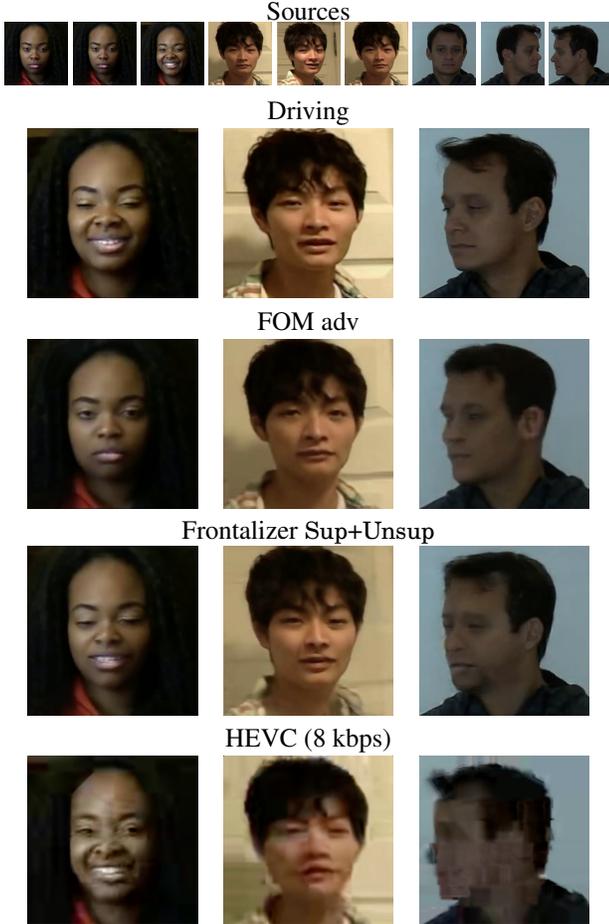


Figure 8. Our mobile Frontalizer models outperform FOM adv, requiring less than 8 kbps, and visibly improve over HEVC.

ity (LPIPS) and landmark fidelity (NME). Expression did not bring improvement in the metrics as detailed in the Appendix. Qualitative comparisons appear in Figure 8. The results show that in low-compute and low-bandwidth regime, it is beneficial to add extra facial landmarks and rely on multiple frames: we observe that eye blinking and large mouth opening are skipped by the FOM mobile baseline.

Influence of SPADE in the decoder. Finally, we train a model combining our Frontalizer approach using the new sampling strategy and SPADE layers in the decoder [11]. The obtained gains are significant both numerically (cf. Table 3) and qualitatively for rendering pupils and teeth. For training this final model, we disabled adversarial training, the learning rate scheduler and let the model train for 400K iterations. The bandwidth of this model is 10.7 kbps at 15 frames per second.

Inference time. Our mobile models (including the landmark tracker) run at 30 FPS on an iPhone 8 using CoreML to access the GPU chip. In our measurements, for mobile models capped at 15 FPS, the battery usage is around 0.2% / min on recent phones (e.g. 2019 iPhone 11 with a specialized Neural Engine chip), and around 0.7% / min on older phones such as the iPhone 8.

4.3. Frame compression

To reconstruct the current frame on the receiver side, we need the source frame(s), and the driving frame’s necessary information. We assume that source frames are sent ahead of time and focus our bandwidth estimate on the driving frame information. A driving sample can be described by a list of floating-point numbers: the average keypoint position, the position of the keypoints relative to this average position, the Jacobians and the expression codes. Rather than sending these values directly, we instead send deltas compared to the previous frame’s values. We furthermore quantize Unsup keypoint coordinates on 12-bits, Sup keypoints on 8 bits, the $4 \times K$ Jacobians on 16 bits in $[-15, 15]$, and the M expression codes on 10 bits in $[-1, 1]$. We then adopt arithmetic coding [13] to compress the data. Arithmetic coding is a lossless entropy coding algorithm, using a prior distribution over the values to encode frequent values using fewer bits. We compute this distribution on the first half of the DFDC val. set, and report compressed sizes on the second half for all models in Table 4.

Model	Bits per frame	Bdth @ 15fps
FOM (Unsup +Jac.)	426	6.4 kbps
Ours (Sup)	264	4.0 kbps
Ours (Sup+Unsup)	418	6.3 kbps
Ours (Sup+Unsup+Expr)	433	6.5 kbps

Table 4. Frame-size for different models, compressed with arithmetic encoding, and measured on 15 fps videos of the test set.



Figure 9. The multiple reference frame FOM baseline still suffers from head deformations and results in non-smooth transitions when dynamically switching the source frame.

4.4. Additional comparisons

We provide numerical comparisons to two competing approaches NTH [21], Hybrid SPADE [11], and a third baseline dubbed ‘‘FOM mult’’ in Table 5. The FOM mult baseline consists in dynamically selecting the reference frame in which landmarks are the closest to the driving one, and apply the FOM approach to reconstruct the result. We include a qualitative comparison to this baseline in Fig. 9 and quantitative results in Table 5. Even if there is a slight drop in LPIPS on dense models, the flickering it induces when the head turns makes it unusable in practice.

	LPIPS ↓	NME ↓	CSIM ↑
FOM [15]	0.116 ▲12%	0.375 ▲31%	0.817 ▲6%
FOM mult	0.092 ▼11%	0.318 ▲19%	0.851 ▲2%
NTH [21]	0.147 ▲31%	0.288 ▲10%	0.785 ▲10%
Ours dense	0.102	0.259	0.864
Hybrid-SPADE [11]	0.166 ▲10%	0.312 ▲9%	0.801 ▲2%
Ours mobile	0.149	0.285	0.821

Table 5. Comparison of dense and mobile methods on DFDC-50. The FOM with multiple reference baseline, reported for reference, is not temporally consistent, thus inapplicable (See Fig. 9).

The Neural Talking Head (NTH) approach [21] suffers from a loss of identity (low CSIM), as illustrated in Fig. 10.

Figure 11 provides a comparison of mobile results with the best model from [11]. Our results are sharper with a twice larger resolution of 256× and preserve expressions better. Most importantly, our approach can be combined with [11], leading to state-of-the-art results on mobile, com-

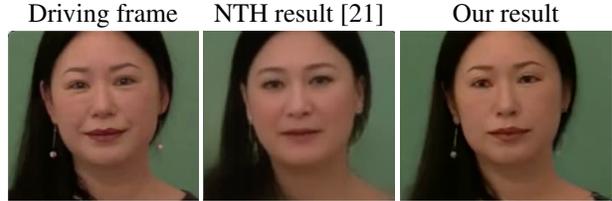


Figure 10. Our Frontalizer preserves identity better than Neural Talking Heads [21].



Figure 11. Comparison to Hybrid Spade [11] on mobile.

binning the strengths of the Frontalizer embedding with the refinement of hybrid SPADE layers.

5. Conclusion

Contributions. We propose a solution to the profile views rendering problem for face animation by designing a frontalized embedding based on a few source images warped to a reference position in the feature space. The resulting multi-view embedding avoids relying on the model to infer unseen areas during inference. We show that our approach compares favorably to baselines, especially in the case of large head rotations, with mobile results outperforming the state-of-the-art by a large margin. We also propose using a lightweight conditioning approach to refine facial expressions improving dense animation approaches.

Limitations. While the majority of examples are correctly rendered with our approaches, there remains some margin for improvements in mobile models for glasses, pupils, head occlusions. Temporal consistency is dependent on the landmark extractor. In general we do not observe problems but slight flickering may happen in case of bad landmark detections.

Broader impacts. Deep facial image generation is a sensitive research topic because reenactment applications with

bad intentions could be harmful. We will not publish the source code for this work, to reduce these risks. We hope that the good performance of our approach in low compute regime will help to unlock video chat in low bandwidth situations.

Acknowledgments

We would like to thank Michael Chang, Patrick Labatut, Jeff LaFlam, Maja Pantic, Thibault Peyronel, Albert Pumarola, Marcy Regalado for their various support to our work.

References

- [1] AI Can See Clearly Now: GANs Take the Jitters Out of Video Calls (Nvidia Maxine blog post), Oct 2020. [2](#)
- [2] 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 *ICCV*, 2017. [6](#)
- [3] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007. [5](#)
- [4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. [6](#)
- [5] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. *ArXiv 2006.07397*, 2019. [5](#), [6](#)
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014. [5](#)
- [7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ArXiv 1412.6980*, 2014. [5](#)
- [8] Goluck Konuko, Giuseppe Valenzise, and Stéphane Lathuilière. Ultra-low bitrate video conferencing using deep image animation. In *ICASSP*, 2020. [1](#), [2](#), [3](#)
- [9] Peirong Liu, Rui Wang, Xuefei Cao, Yipin Zhou, Ashish Shah, Maxime Oquab, Camille Couprie, and Ser-Nam Lim. Self-appearance-aided differential evolution for motion transfer. In *Arxiv*, volume 2110.04658, 2021. [2](#)
- [10] Arsha Nagrani, Joon Son. Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Interspeech*, 2017. [6](#)
- [11] Maxime Oquab, Pierre Stock, Oran Gafni, Daniel Haziza, Tao Xu, Peizhao Zhang, Onur Celebi, Yana Hasson, Patrick Labatut, Bobo Bose-Kolanu, Thibault Peyronel, and Camille Couprie. Low bandwidth video-chat compression using deep generative models. In *CVPR Mobile AI Workshop*, 2020. [1](#), [2](#), [3](#), [6](#), [8](#), [9](#)
- [12] Albert Pumarola, Antonio Agudo, Aleix M. Martínez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, 2018. [2](#)
- [13] Jorma Rissanen and Glen G Langdon. Arithmetic coding. *IBM Journal of research and development*, 23(2):149–162, 1979. [8](#)
- [14] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. [6](#), [10](#)
- [15] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. [1](#), [2](#), [3](#), [4](#), [5](#), [7](#), [9](#), [10](#), [11](#)
- [16] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021. [5](#), [6](#)
- [17] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. [1](#), [2](#), [3](#)
- [18] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018. [2](#)
- [19] Hui Yu, Oliver GB Garrod, and Philippe G Schyns. Perception-driven facial expression synthesis. *Computers & Graphics*, 36(3):152–162, 2012. [5](#)
- [20] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. *ECCV*, 2020. [2](#)
- [21] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *ICCV*, 2019. [2](#), [9](#)
- [22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [6](#)
- [23] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *CVPR*, pages 192–201, 2017. [5](#)
- [24] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021. [2](#)

6. Appendix

6.1. Architectures and training details

6.1.1 Frontalizer

The Frontalizer module contains an hourglass network and a set of keypoints defining the reference position p_r . In the server (*dense*) models, the architecture of the hourglass net is identical to the one of the Dense Motion network of [15]. In this section we will focus on describing the mobile architectures.

The base building block of our networks is the inverted residual block introduced in [14], that we describe in Figure 13. We then describe the mobile architectures for the

different networks: the Frontalizer, Dense Motion, Keypoint and expression networks. All of them use the hour-glass architecture defined in Figure 15.

The number of source frames used for training is set to $N = 2$ (N can be higher during inference). We note K the number of unsupervised keypoints when applicable, and we note n_{kp} the total number of keypoints including the facial landmarks.

The Frontalizer network takes as input a tensor of size $(N, n_{kp}, 32, 32)$, where each map of size 32×32 describes a keypoint displacement as a difference of Gaussians. It outputs N flow maps of size $(32, 32, 2)$ and N confidence maps of size $(32, 32)$. The flow maps are then applied to the encoder feature for each of the N images to obtain warped features, and the confidence maps are softmax-ed, then used for obtaining the weighted sum of the warped features, corresponding to the multi-view embedding.

The dense motion network takes as input a tensor of size $(n_{kp}, 32, 32)$ describing the keypoints displacements, and outputs a one flow map of size $(32, 32, 2)$ and one confidence map of size $(32, 32)$, that are both applied to the multi-view embedding.

The (unsup) keypoint extractor takes as input an image tensor of size $(3, 64, 64)$ and outputs K heatmaps of size $(64, 64)$. We apply a spatial softmax to these heatmaps then compute the center of mass of the resulting maps to output $2 \times K$ keypoint coordinates.

The Encoder, described in Figure 14, takes as input an image tensor of size $(3, 256, 256)$ and outputs a feature map of size $(32, 32, 32)$. The Decoder, described in Figure 16, takes as input a feature map of size $(32, 32, 32)$ and outputs an image of size $(3, 256, 256)$.

6.1.2 Dense expression module

The MLP g is a two-layer network taking in input a 4096 dimensional vector for dense models, or 512 dimensional for mobile models. It consists of a linear layer mapping to a 32 dimensional layer, a batch normalization layer, a ReLU, a linear layer mapping to a 16 dimensional vector and a Sigmoid. For training, the weight γ_E of the expression loss is set to 40.

6.1.3 Training losses

In addition to the multi-scale perceptual loss, and equivariance loss in settings employing Unsup keypoints, we employ a feature matching loss, and an adversarial loss. We use the same hyper-parameters as defined in [15] to weight the losses. Differently to [15], we turned adversarial training on from the beginning of training for Frontalizer models. Indeed, while for the FOM baseline adversarial fine-tuning was experimentally better than direct training in the

dense setting, for our other models we performed adversarial training from scratch.

Additional losses for Sup+Unsup models. We tried to add Jacobians as in [15] but it did not change performance, cf. Table 6.

6.2. Additional results

6.2.1 Cross-reenactment

We finally show that our approach is able to provide cross-reenactment results as illustrated in Figure 12.

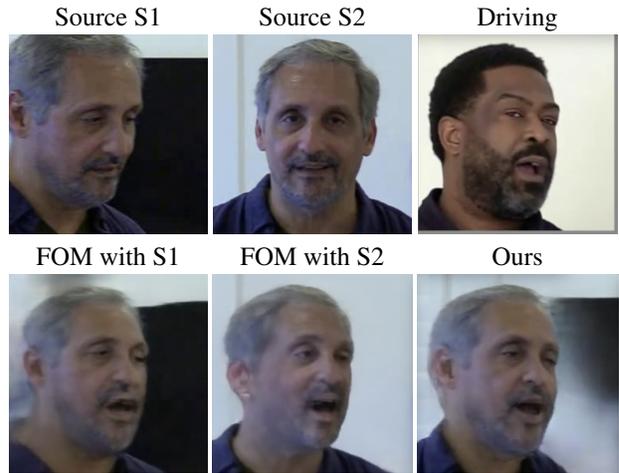


Figure 12. Cross reenactment: Face animation given a few source frames using landmarks from a driving video.

6.2.2 Head shape and expression preservation

Figure 17 provides examples of improvements of the Frontalizer Sup+Unsup model over the Frontalizer Sup dense model.

Figure 18 shows examples of additional improvements brought by expression conditioning on top of the Frontalizer Sup+Unsup dense model.

6.2.3 Ablation study on mobile models

Metrics using expression in mobile models are provided in Table 7. From this table, it seems that expression conditioning does not bring much to the Sup+ Unsup Frontalizer. Qualitatively, we observe improvement of eye rendering, see Figure 19.

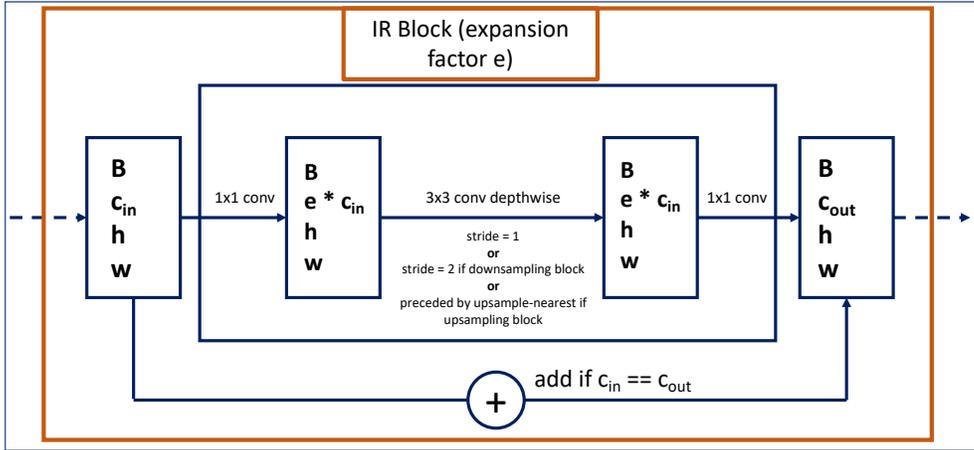


Figure 13. MobileNetV2 Inverted Residual block with expansion factor e , for a number of input channels c_{in} and output channels c_{out} .

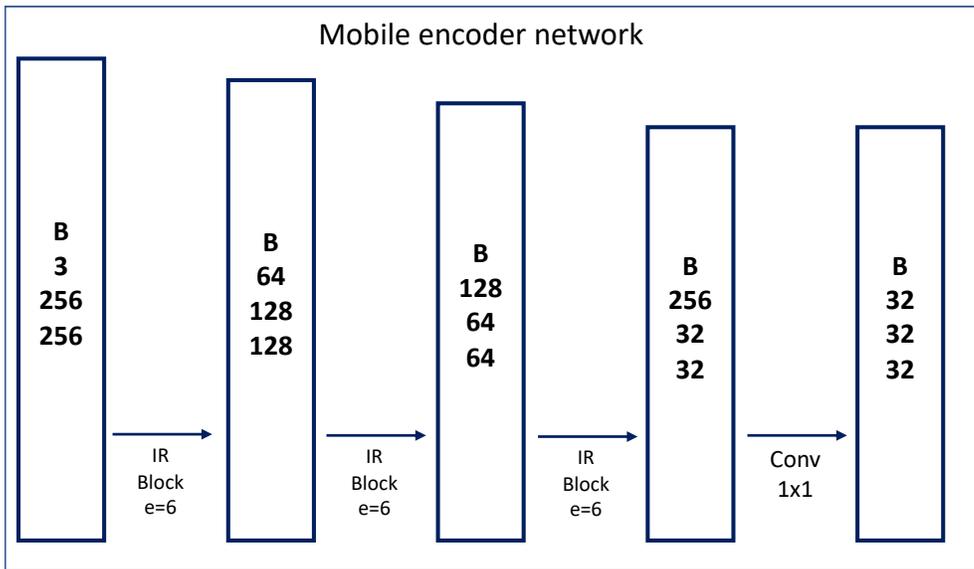


Figure 14. Mobile encoder architecture used in our mobile setups

	DFDC-rotations				VoxCeleb			
	LPIPS ↓	NME ↓	CSIM ↑	Expr ↑	LPIPS ↓	NME ↓	CSIM ↑	Expr ↑
Frontalizer sup+unsup + free expr with Jacobians	0.113 ±0.03	0.467 ±0.38	0.789 ±0.06	0.897 ±0.05	0.140 ±0.04	0.256 ±0.04	0.781 ±0.06	0.910 ±0.05
Frontalizer sup+unsup + free expr, no Jacobians	0.115 ±0.03	0.463 ±0.42	0.787 ±0.06	0.894 ±0.05	0.147 ±0.04	0.247 ±0.04	0.779 ±0.06	0.909 ±0.04

Table 6. Ablation study on the impact of Jacobians in the dense Sup Unsup Expr model.

	DFDC-rotations				VoxCeleb			
	LPIPS ↓	NME ↓	CSIM ↑	Expr ↑	LPIPS ↓	NME ↓	CSIM ↑	Expr ↑
Ours (sup+unsup)	0.148 ±0.003	0.585 ±0.044	0.723 ±0.006	0.865 ±0.005	0.165 ±0.002	0.275 ±0.003	0.747 ±0.003	0.893 ±0.002
Ours (sup+unsup+expr)	0.151 ±0.003	0.565 ±0.044	0.718 ±0.006	0.865 ±0.005	0.167 ±0.002	0.273 ±0.003	0.748 ±0.003	0.886 ±0.002

Table 7. Ablation study on the impact of expressions in the mobile Sup Unsup model (no Jacobians).

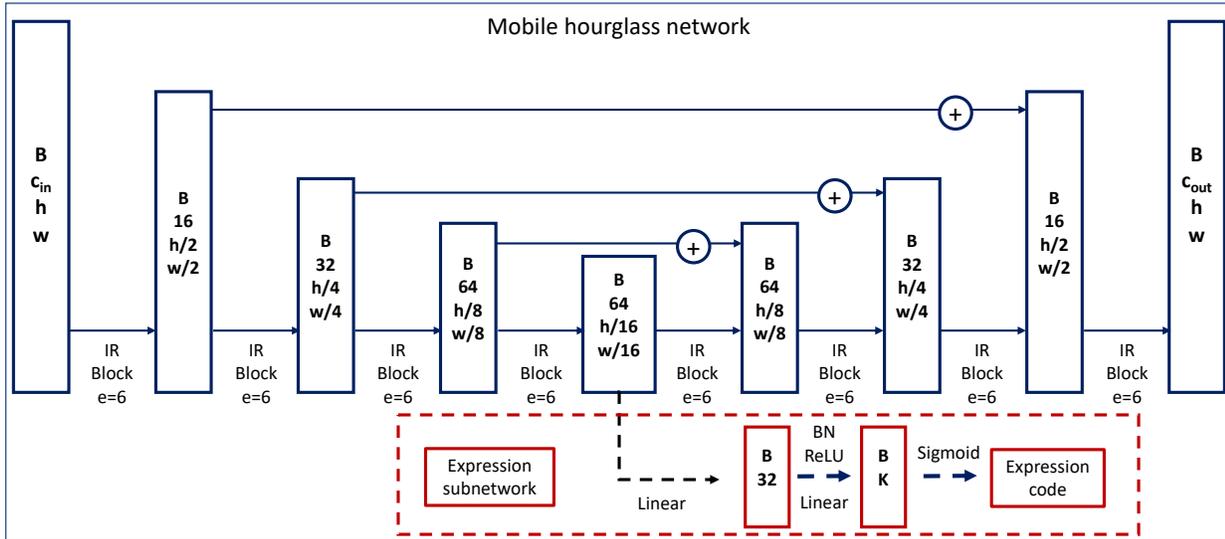


Figure 15. Mobile hourglass architecture used in our mobile setups, for the Keypoint Detector, Dense Motion and Frontalizer networks. In order to extract the expression code, we add a small MLP network that uses the central feature of the Keypoint Detector network.

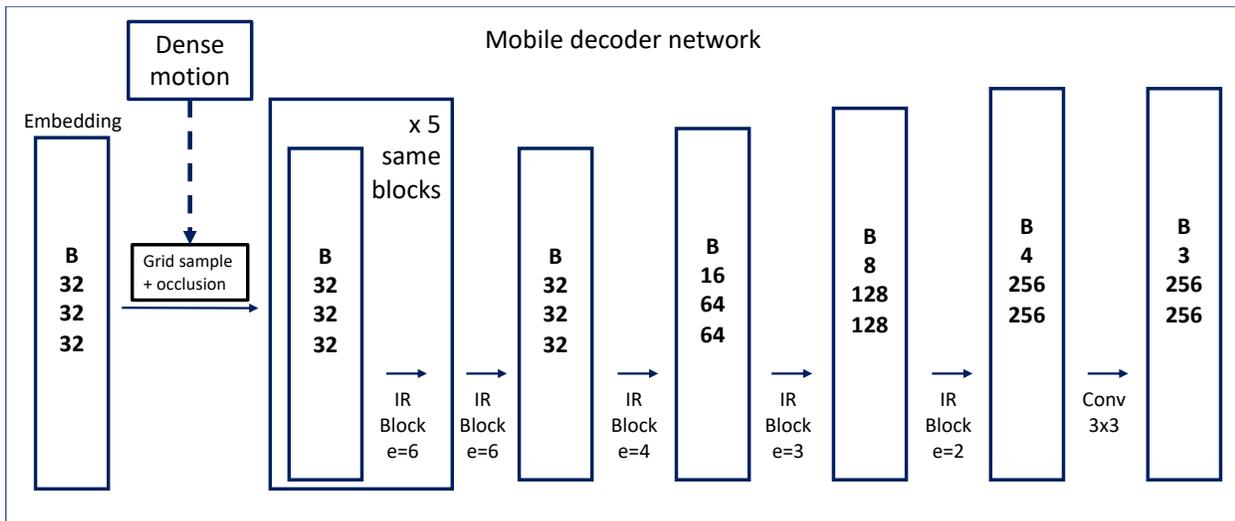


Figure 16. Mobile decoder architecture used in our mobile setups



Figure 17. Head shape improvements on top of Frontalizer sup. Left: original; Middle Frontalizer sup. Right: Frontalizer sup+unsup.



Figure 18. Expression improvements on top of Frontalizer un-sup+sup. Left: original; Middle Frontalizer baseline; right: baseline with free expression conditioning.

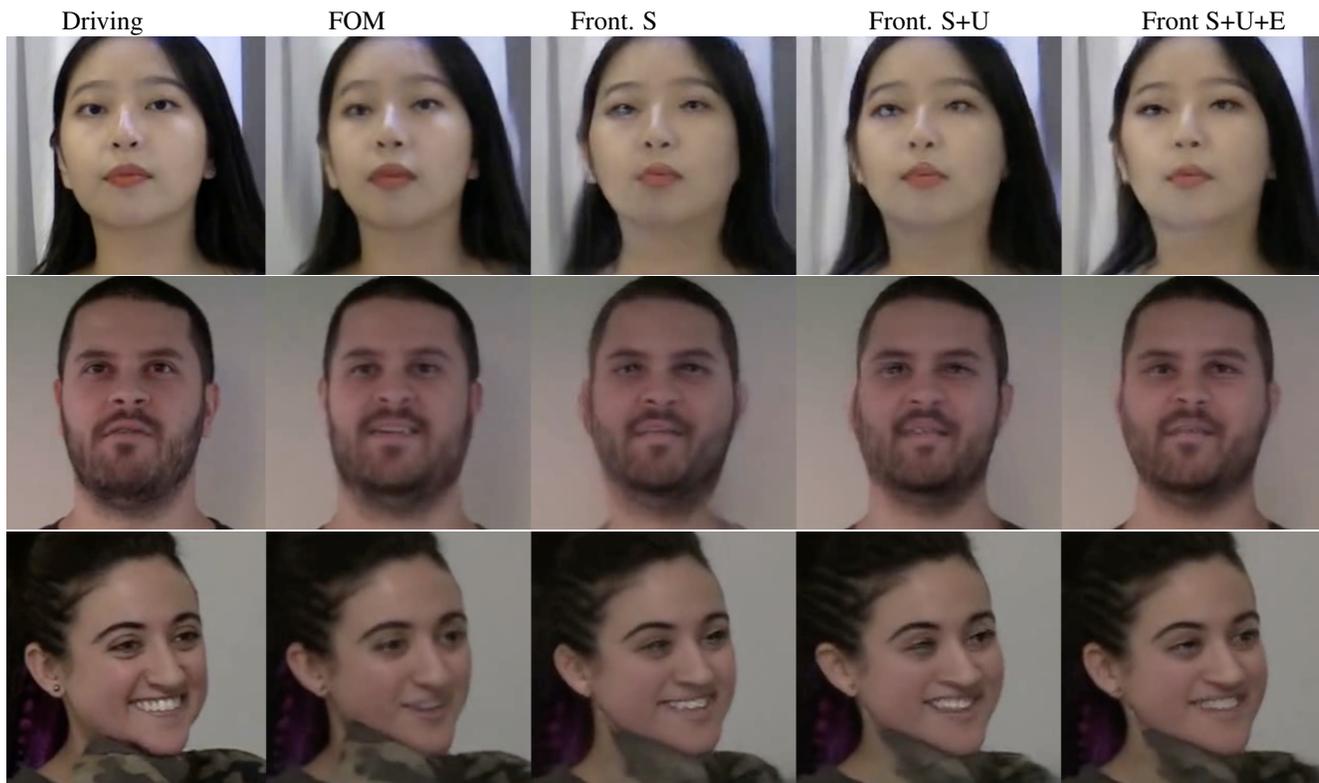


Figure 19. Mobile ablation results. Expression conditioning improves eye rendering.