Expression-preserving face frontalization improves visually assisted speech processing

Zhiqi Kang, Mostafa Sadeghi, Radu Horaud, Xavier Alameda-Pineda

International Journal of Computer Vision, 2023

Expression-preserving face frontalization improves visually assisted speech processing

Zhiqi Kang · Mostafa Sadeghi · Radu Horaud · Xavier Alameda-Pineda

Abstract Face frontalization consists of synthesizing a frontal view from a profile one. This paper proposes a frontalization method that preserves non-rigid facial deformations, i.e. facial expressions. It is shown that expression-preserving frontalization boosts the performance of visually assisted speech processing. The method alternates between the estimation of (i) the rigid transformation (scale, rotation, and translation) and (ii) the non-rigid deformation between an arbitrarily-viewed face and a face model. The method has two important merits: it can deal with non-Gaussian errors in the data and it incorporates a dynamical face deformation model. For that purpose, we use the Student's t-distribution in combination with a Bayesian filter in order to account for both rigid head motions and time-varying facial deformations, e.g. caused by speech production. The zero-mean normalized cross-correlation (ZNCC) score is used to evaluate the ability of the method to preserve facial expressions. The method is thoroughly evaluated and compared with several state of the art methods, either based on traditional geometric models or on deep learning. Moreover, we show that the method, when incorporated into speech processing pipelines, improves word recognition rates and speech intelligibility scores by a considerable margin.[1]

This work has been partially supported by the H2020 SPRING project #871245 and by the Multidisciplinary Institute of Artificial Intelligence (MIAI) ANR-19-P3IA-0003.

Z. Kang, R. Horaud, X. Alameda-Pineda
Inria Grenoble & Université Grenoble Alpes, France

M. Sadeghi
Inria Nancy Grand-Est
Villers-lès-Nancy, France

[1] Supplemental material is accessible at https://team.inria.fr/robotlearn/research/facefrontalization.

Keywords face frontalization · Student's t-distribuyion · robust point registration · Bayesian filtering · lip reading · audio-visual speech enhancement · variational auto-encoders.

1 Introduction

Face frontalization is the problem of synthesizing a frontal view of a face from an arbitrarily viewed one. Recent research has shown that face frontalization consistently boosts the performance of face recognition, e.g. Yim et al (2015); Zhu et al (2015); Banerjee et al (2018); Zhao et al (2018); Zhou et al (2018, 2020). It is worth noticing that face recognition requires *expression-free* face frontalization (which is also referred to as face normalization). In contrast, other applications, such as facial expression recognition, e.g. Pei et al (2020) and visual speech processing, e.g. Fernandez-Lopez and Sukno (2018); Adeel et al (2019); Martinez et al (2020); Cheng et al (2020), require *expression-preserving* face frontalization. In this paper we present a novel face frontalization methodology that combines robust statistical inference with a dynamic model. We show that the proposed algorithms improve the performance of visual speech by a considerable margin.

It has long been established that visual perception plays a primordial role in speech communication. In particular, vision provides an alternative representation of some of the information that is present in the audio, with the advantage that it is affected neither by acoustic noise nor by competing audio sources. The most prominent visual features used in human-to-human, human-to-computer and human-to-robot interactions are facial movements. Facial movements are a combination

of rigid head movements and non-rigid facial deformations. On one side, head movements play linguistic functions as they mark the structure of the ongoing discourse and are used to regulate interaction McClave (2000). On the other side, lip and jaw movements are generated by facial muscles which, in turn, are controlled by speech production – they are correlated with phonemes and with word pronunciation Schultz et al (2017). Hence visual information plays a fundamental function both in speech recognition and in speech intelligibility.

In particular, automatic speech recognition (ASR) and speech enhancement (SE) play crucial yet complimentary roles in speech communication systems. SE aims to improve the quality of noisy speech signals to be used by ASR. It is well established that audio speech enhancement (ASE) is severely limited in adverse acoustic situations, e.g. background noise. Multimodal speech enhancement, and in particular audio-visual speech enhancement (AVSE) aims at incorporating the complimentary information available with visual information. Lip reading plays a similar role in ASR.

AVSE has received a lot of attention in the recent past, mainly because of the advent of deep neural networks (DNNs) which have considerably boosted their performance Michelsanti et al (2021). Nevertheless, the vast majority of existing methods assume clean visual information – they take as input lip regions that are cropped from frontal and steady face images, Hou et al (2018); Sadeghi et al (2020); Adeel et al (2021). Currently there are no DNN architectures able to mitigate the effect of rigid head motions that are inherently present in speech communication. Moreover, the vast majority of existing datasets for training and testing AVSE are recorded in constrained conditions – the participants were instructed to avoid head movements and to face the camera Abdelaziz (2017); Anina et al (2015). As for lip reading Fernandez-Lopez and Sukno (2018), although there were some attempts to deal with in the wild datasets, the current state of the art is limited to the task of isolated word recognition (IWR) Chung and Zisserman (2016); Ma et al (2021a). Not surprisingly, the performance of existing methods rapidly degrades in the presence of noisy visual information. Therefore, although these methods have profited from state of the art deep-learning models, they are ineffective in realistic conversational scenarios.

In this paper we are interested in investigating vision-assisted speech processing methods that are robust with respect to noisy lip movements caused by head motions, e.g. Figure 1 and Figure 2. We propose to incorporate face frontalization (FF) into visual and audio-visual

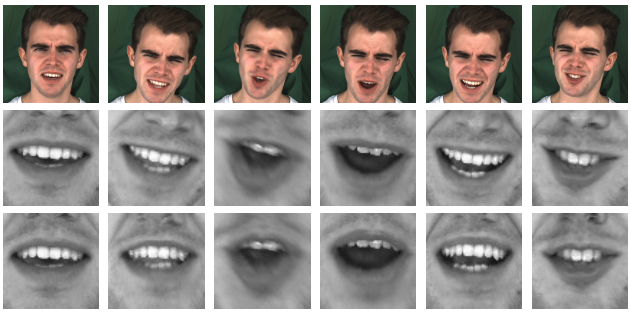
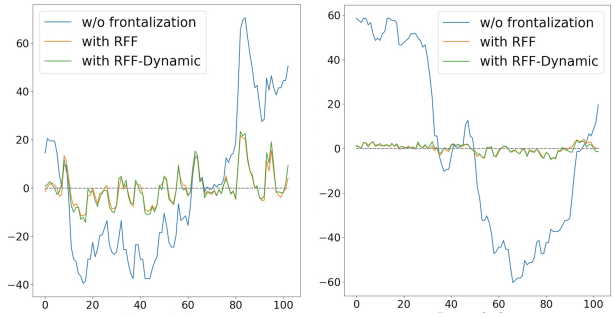


Fig. 1: An example of applying expression-preserving face frontalization to a person that utters speech. Top: input images; Middle: lip regions before removing head movements; Bottom: lip regions after removing head movements.



(a) Vertical lip motion (b) Horizontal lip motion

Fig. 2: Lip motion without frontalization (blue), with robust frontalization (orange) Kang et al (2021), and with robust-dynamic frontalization (green), proposed in this paper. The curves correspond to the example of Figure 1. The curves correspond to image-plane displacements in pixels as a function of the frame number and they show the motion of the landmark located at the center of the upper lip, for the example shown on Figure 1.

speech processing DNN pipelines. Visual speech processing necessitates a FF method that guarantees that non-rigid facial deformations are preserved. Moreover and unlike FF for face recognition from a single image, FF for visual speech analysis must incorporate a dynamic model in order to capture the temporal nature of lip movements. We address these challenging problems on the following grounds: (i) the image warping needed by FF should be guided by a rigid transformation, (ii) the estimation of this transformation should be robust with respect to non-rigid deformations, and (iii) a dynamic face deformation model is needed in order to characterize the temporal behaviour of lip and jaw movements associated with speech.

The rationale of the proposed method is to decouple rigid head pose from non-rigid facial deformations. These two pieces of information are encoded in the observed 3D landmarks while they are affected by two types of errors: small detection errors that are indistinguishable from non-rigid deformations, and large localization errors that might strongly biais the results. For these reasons, the estimation problem at hand is cast into the problem of robust statistical inference. Head pose and non-rigid deformation are coupled with the *generalized Student's t-distribution* Forbes and Wraith (2014) – a heavy tailed probability distribution function (pdf) that is able to deal both with Gaussian inliers and with non-Gaussian outliers. The associated expectation conditional maximization (ECM) procedure alternates between (i) the evaluation of the posterior distributions of weights associated with observed facial landmarks, (ii) the estimation of the rigid head-pose parameters (scale, rotation and translation) and (iii) the estimation of the parameters of a deformable face model, e.g. Blanz and Vetter (1999), i.e. Figure 3-(a). The landmark weights just mentioned have a ponderable role: the higher is the weight, the more reliable is the landmark.

In the past the Student's t-mixture model (TMM) Peel and McLachlan (2000) was used for the task of robust non-rigid registration of multiple point sets Zhou et al (2014); Ravikumar et al (2018). These methods jointly register the points and estimate the rigid transformations that allow to optimally align the sets, on the premise that a majority of points in the sets are in rigid correspondence. In the case of landmark-based FF both rigid and non-rigid alignment are needed while it is not necessary to perform registration, hence a single pdf, and not a mixture, is sufficient.

We also propose a dynamical extension. The frontalized landmarks are treated as observations of a linear dynamical system (LDS). Unlike a standard LDS, the proposed one is equipped with two sequences of latent variables governed by two interconnected linear-Gaussian dynamical regimes, i.e. Figure 3-(b). The two latent variables correspond to the 3D vertex coordinates and to the low-dimensional face embedding of a 3D morphable model (3DMM), respectively. The 3D vertices at the current frame are stochastically generated from the 3D vertices at the previous frame. Similarly, the current shape embedding is stochastically generated from the previous embedding. At each frame, the vertices are reconstructed from the shape embedding. In turn, the vertices stochastically generate the frontalized landmarks. We provide a formal derivation of the proposed *doubly latent* LDS and we show that it can be reformulated as a standard Kalman filter, with its associated recursive solver.

We empirically evaluate the performance of FF using the zero-mean normalized cross correlation (ZNCC) score, Sun (2002), between a frontalized face and its ground-truth frontal counterpart. We embed FF into a state of the art deep lip reading model Ma et al (2021a). We also use FF in combination with a recently proposed deep AVSE model Sadeghi et al (2020); Sadeghi and Alameda-Pineda (2021). We use three different datasets associated with these three sets of experiments and we compare our method with two traditional frontalization methods Hassner et al (2015), Banerjee et al (2018), and with two methods based on generative adversarial networks (GANs), Zhou et al (2020), Yin et al (2020). We show that the proposed expression-preserving face frontalization method outperforms all the other methods – either based on traditional computer vision or based on DNNs – by a considerable margin. A prominent result is that robust estimation of the rigid transformation underlying FF outperforms GAN-based frontalization. Indeed, the latter estimates millions of parameters of a non-linear image-to-image mapping, which cannot guarantee that the non-rigid facial deformations, e.g. expressions and lip movements, are preserved.

The remainder of this article is organized as follows. Section 2 summarizes the related work. Section 3 describes in detail the proposed expression-preserving frontalization framework. Section 4 provides algorithm implementation details. Section 5 describes a benchmark based on the ZNCC score. Section 6 describes experiments with a lip-reading dataset. Section 7 combines face frontalization with audio-visual speech enhancement. Finally Section 8 draws some conclusions.

2 Related Work

As already mentioned, face frontalization consists of synthesizing a frontally-viewed face from an arbitrarily-viewed face. Recently, a successful approach has been to train DNNs in order to learn a non-linear 2D-to-2D mapping between an arbitrary view and a frontal view. Some of the best performing DNN-based frontalization methods use CNN-GAN architectures, e.g. Yin et al (2017); Huang et al (2017); Tran et al (2017); Zhao et al (2018); Zhang et al (2019); Rong et al (2020); Zhang et al (2021); Yin et al (2020), which outperform CNN-only models, e.g. Yim et al (2015). These methods necessitate large collections of input/output pairs of face images. For that purpose Zhang et al (2019);

Yin et al (2020); Zhang et al (2021) use two datasets that contain multiple-camera recordings in a controlled setup, i.e. Gao et al (2007); Gross et al (2010). Zhang et al (2019) proposed to learn dense pixel-to-pixel correspondences between the input-output faces. Subsequently, Zhang et al (2021) proposed a semi-supervised GAN-based method that augments the paired face images of Gross et al (2010) with unpaired in-the-wild faces with large variations in identity, e.g. Huang et al (2008); their adversarial and identity-preserving losses enhance face recognition performance. Yin et al (2020) proposed a dual-attention GAN architecture that captures long-term dependencies in image space, thus providing a mean to preserve identity. These DNN-based methods are designed to predict as-neutral-as-possible frontal faces, i.e. expression-free faces, in order to improve the performance of face recognition. On the one side, the profile/frontal pairs of Gao et al (2007); Gross et al (2010) are collected in controlled settings in terms of illumination and expression. On the other side, the non-frontal images from in the wild datasets do not have their frontal counterparts to allow frontalization training.

Another way to estimate the non-linear 2D-to-2D mapping between a profile image and a canonical image of a face is to use a rectification network that learns local homographies between a deformed grid, that corresponds to a profile view, and a regular grid, that supposedly corresponds to a frontal view Zhou et al (2018). While this method is well suited for improving the performance of face recognition, it is unable to take into consideration off-the-image-plane rotations, to guarantee a frontal image and to separate rigid head pose from non-rigid facial deformations.

Other methods estimate the pose of an input face with respect to a frontal 3D face model, then use the pose parameters to warp the facial pixels from the input image onto a frontal one. These methods capitalize on pose estimation from 2D-to-3D point correspondences, e.g. Zhu et al (2015); Hassner et al (2015); Ferrari et al (2016); Banerjee et al (2018). In Hassner et al (2015) it was proposed to use a 3D generic model of a face from which a frontal face is generated: 48 facial landmarks (2D) are extracted from the input face and from the neutral and frontal face model (3D), thus providing 2D-to-3D correspondences between the input face and the generic 3D model. This amounts to estimate the intrinsic camera parameters as well as the rigid pose. Similar methods were proposed by Zhu et al (2015), Ferrari et al (2016) and Banerjee et al (2018). Note that with this setup there is an inherent large discrepancy between the expressive input face and the neutral model face. Hence, these methods lack a built-in robust statistical model that enables accurate inference in the presence of large errors in landmark localization and of non-rigid facial deformations. To mitigate this issue, Hassner et al (2015) manually removes jaw landmarks and Zhu et al (2015) purposely removes expressions in order to favour identity features.

Recently, Zhou et al (2020) proposed to synthesize profile views from a collection of frontal views in order to create input/target pairs for the purpose of training image-to-image translation GANs. Their method starts by fitting a 3D face model to a frontal view, using the 2D-to-3D alignment technique of Zhu et al (2019), followed by rotating and rendering the fitted 3D model to obtain a profile view, and finally rotating and rendering it back to reconstruct a frontal view. To summarize, Zhou et al (2020) uses Zhu et al (2019) to estimate the rigid pose and the face deformation parameters in order to frontalize the face, and Zhu et al (2017) to fill in the occluded regions caused by frontalization. Although this method yields state-of-the-art results for the task of face recognition, there is no guarantee that non-rigid facial deformations are preserved by the profile-to-frontal mapping process.

Interestingly, there has only been a handful of attempts to combine dynamic models with facial shape deformation. Baumberg (1998); Lee et al (2007) use a Kalman filter to track a face in an image sequence and to initialize the parameters of a deformable shape model. In Prabhu et al (2010) a Kalman filter is used to predict the location of individual landmarks, from the previous frame to the current frame, and to use these predictions to initialize the parameters of a deformable shape model. The dynamical model that we propose in this paper is totally different because it dynamically updates a deformable model with two interconnected latent variables – this dynamic face-deformation model is inferred in alternance with landmark frontalization.

The proposed method requires 3D facial landmarks. Recently there has been a flourishing literature on this topic, yielding several DNN 3D face alignment (3DFA) models and associated software packages, e.g. Bulat and Tzimiropoulos (2016); Zhu et al (2016); Feng et al (2018); Deng et al (2018); Zhu et al (2019); Jiang et al (2019); Tu et al (2020); Ning et al (2020). We thoroughly analysed and benchmarked four publicly available 3DFA software packages. The results reported in this paper were obtained with the method of Bulat and Tzimiropoulos (2016). The latter is trained using a very large dataset Bulat and Tzimiropoulos (2017) and it assumes a weak-perspective camera model. Recently, it has been shown that the perspective camera

model is better suited for guaranteeing the separation of rigid and non-rigid facial deformations Sariyanidi et al (2020). We propose an alternative rigid/non-rigid separation formulation based on robust statistics.

This article is an extended version of Kang et al (2021) and it contains two extensions: a dynamic face-deformation model and its inference based on linear dynamical systems, Section 3.3, and an in depth investigation of the effect of face frontalization on the performance of audio-visual speech enhancement, Section 7.

3 Expression-preserving face frontalization

In this section we describe in detail the static and dynamic models that reside at the core of the proposed expression-preserving FF framework. These models are graphically represented in Figure 3. After briefly presenting the face deformation model, we describe in detail the estimation of the head-pose and face-deformation parameters, followed by a description of the dynamical formulation and the associated statistical inference. The final stage consist of warping the input face onto a frontal view in such a way that facial deformations remain invariant.

3.1 Face deformation model

In order to model non-rigid facial deformations, we consider a 3D deformable shape model Blanz and Vetter (1999). Such a model is learnt from a training set of 3D faces, or meshes, $\mathcal{M} = \{\boldsymbol{M}_m\}_{m=1}^M$. Each face m in the training set is described by N 3D vertices, namely $\boldsymbol{M}_m = (\boldsymbol{M}_{m1}, \dots, \boldsymbol{M}_{mn}, \dots, \boldsymbol{M}_{mN})^\top \in \mathbb{R}^{3N}$; moreover, the faces are registered: their vertices are in one-to-one correspondence. Let $\mathbf{C} = 1/M \sum_{m=1}^M (\boldsymbol{M}_m - \overline{\boldsymbol{M}})(\boldsymbol{M}_m - \overline{\boldsymbol{M}})^\top$ be the covariance matrix associated with this training set, where the *mean shape* is defined by $\overline{\boldsymbol{M}} = 1/M \sum_{m=1}^M \boldsymbol{M}_m$, and let $(\boldsymbol{\Lambda}, \mathbf{U})$ be the K principal eigenvalue-eigenvector pairs of $\mathbf{C}$, where $\boldsymbol{\Lambda} = \text{Diag}\left(\lambda_1, \dots, \lambda_K\right)$, with $\lambda_1 \geq \dots \geq \lambda_K \geq 0$ and $\mathbf{U} = \left(\boldsymbol{U}_1 \dots \boldsymbol{U}_K\right) \in \mathbb{R}^{3N \times K}$ is a column-orthogonal matrix, i.e. $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_K$, with $K \ll 3N$.

The vertices of a face $\boldsymbol{V} = (\boldsymbol{V}_1, \dots, \boldsymbol{V}_n, \dots, \boldsymbol{V}_N)^\top \in \mathbb{R}^{3N}$ can be projected onto the low-dimensional space spanned by the principal eigenvectors, namely

$$\boldsymbol{s} = \mathbf{U}^\top (\boldsymbol{V} - \overline{\boldsymbol{M}}), \tag{1}$$

where $\boldsymbol{s} \in \mathbb{R}^K$ is the face embedding (or encoding). Conversely, it is possible to reconstruct (or decode) the face $\boldsymbol{V}$ from its embedding $\boldsymbol{s}$, and this up to a *decoding* error $\boldsymbol{F}$:

$$\boldsymbol{V} = \mathbf{U}\boldsymbol{s} + \overline{\boldsymbol{M}} + \boldsymbol{F}, \text{ s.t. } \boldsymbol{s}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{s} \leq 1. \tag{2}$$

The above inequality constrains the reconstructed mesh to correspond to an embedding $\boldsymbol{s}$ that lies inside an ellipsoid with half axes equal to $\sqrt{\lambda_k}$. This guarantees with 99% confidence that $\boldsymbol{V}$ belongs to the space spanned by the training set. Therefore, each vertex $\boldsymbol{V}_n$ of $\boldsymbol{V}$ can be reconstructed from $\boldsymbol{s}$ with

$$\boldsymbol{V}_n = \mathbf{U}_n \boldsymbol{s} + \overline{\boldsymbol{M}}_n + \boldsymbol{F}_n = \mathbf{W}_n \boldsymbol{S} + \boldsymbol{F}_n, \tag{3}$$

where $\mathbf{U}_n \in \mathbb{R}^{3 \times K}$ is such that $\mathbf{U} = (\mathbf{U}_1 \dots \mathbf{U}_n \dots \mathbf{U}_N)$, $\boldsymbol{S} = [\boldsymbol{s}; 1] \in \mathbb{R}^{K+1}$ (vertical concatenation) and $\mathbf{W}_n = (\mathbf{U}_n \; \overline{\boldsymbol{M}}_n) \in \mathbb{R}^{3 \times (K+1)}$.

3.2 Head-pose and face-deformation estimation

We now consider an image of a face with an unknown pose. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_j, \dots, \boldsymbol{X}_J)^\top \in \mathbb{R}^{3J}$ be a vector of J 3D landmarks extracted from this face. The pose is parameterized by a rigid transformation, namely scale $\rho \in \mathbb{R}^+$, rotation $\mathbf{R} \in \text{SO}(3)$, and translation $\boldsymbol{T} \in \mathbb{R}^3$, between the observed landmarks $\boldsymbol{X}$ and the frontalized landmarks $\boldsymbol{Y} \in \mathbb{R}^{3J}$:

$$\boldsymbol{Y}_j = \rho \mathbf{R} \boldsymbol{X}_j + \boldsymbol{T}, \; \forall j \in \{1 \dots J\}. \tag{4}$$

These landmarks correspond to J vertices, annotated such that there is a one-to-one correspondence between $\{\boldsymbol{V}_j\}_{j=1}^J$ and $\{\boldsymbol{Y}_j\}_{j=1}^J$ up to an error $\boldsymbol{D}_j$. We have:

$$\boldsymbol{Y}_j = \boldsymbol{V}_j + \boldsymbol{D}_j = \mathbf{W}_j \boldsymbol{S} + \boldsymbol{E}_j, \tag{5}$$

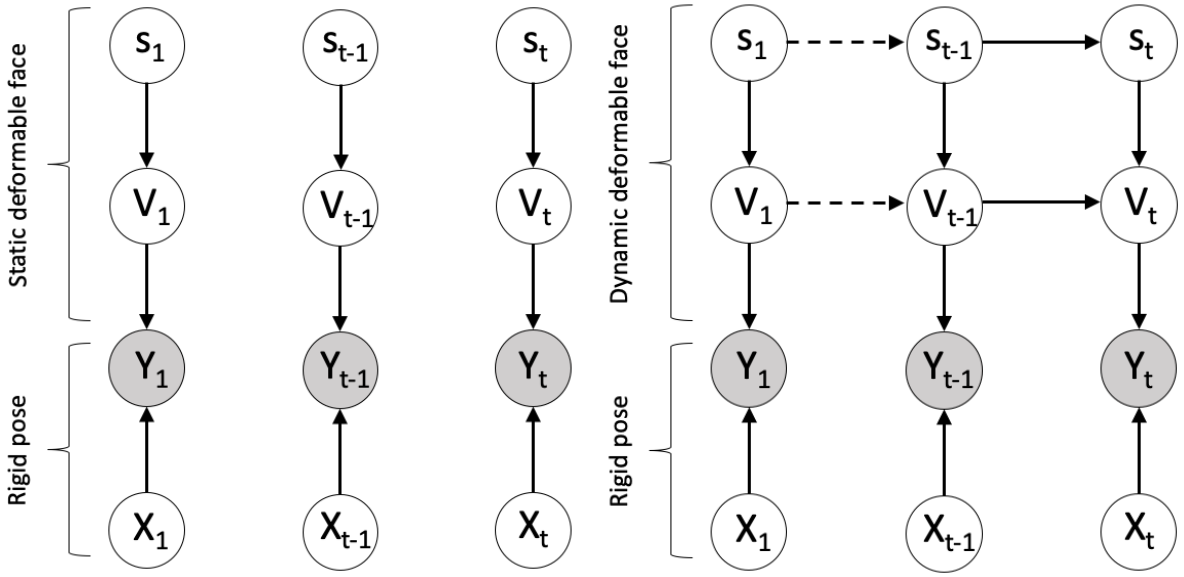
where $\boldsymbol{E}_j = \boldsymbol{D}_j + \boldsymbol{F}_j$ is the total error. By combining the above equations, we obtain:

$$\boldsymbol{E}_j = \rho \mathbf{R} \boldsymbol{X}_j + \boldsymbol{T} - (\mathbf{U}_j \boldsymbol{s} + \overline{\boldsymbol{M}}_j), \tag{6}$$

Assuming that these error vectors are random variables drawn from a probability distribution function (pdf) we can write a maximum likelihood estimator (MLE), or equivalently, the minimization of the following negative log-likelihood function:

$$\mathcal{L}(\boldsymbol{\theta} | \boldsymbol{X}) = -\sum_{j=1}^J \log p(\boldsymbol{E}_j; \boldsymbol{\theta}), \tag{7}$$

where $\boldsymbol{\theta}$ is the vector of model parameters, i.e. the rigid parameters, the shape parameters and the pdf parameters. From (6) one may see that it is possible to alternate between the estimation of the shape parameters $\boldsymbol{s}$ and the rigid (frontalization) parameters ρ, $\mathbf{R}$, and $\boldsymbol{T}$.



(a) Head-pose and face-deformation estimation (b) Dynamic face-deformation inference

Fig. 3: Both models alternate between the estimation of the rigid head-pose parameters and the estimation of the face-deformation parameters. The arrows illustrate stochastic generative models. In both cases, the frontalized landmarks $\boldsymbol{Y}$ are generated from the observed landmarks $\boldsymbol{X}$. The static model (a) uses the Student's t-distribution to estimate the rigid (head pose) and non rigid (face deformation) parameters. In addition, the temporal model (b) makes use of two stochastic dynamical regimes, one that governs the evolution of the shape embedding $\boldsymbol{s}$ and a second one that governs the evolution of the shape vertices $\boldsymbol{V}$. Note that these two latent variables are interconnected and that the frontalized landmarks are used as observations by the proposed doubly-latent LDS.

Because the 3D landmarks are affected by noise and by non-rigid facial deformation, we opt for a robust pdf, namely the generalized Student's t-distribution:

$$p(\boldsymbol{E}_j;\boldsymbol{\theta}) = \int_0^\infty \mathcal{N}(\boldsymbol{E}_j;0,\omega_j^{-1}\boldsymbol{\sigma})\mathcal{G}(\omega_j;\mu,1)d\omega_j, \tag{8}$$

where $\mathcal{N}(\boldsymbol{E};\mathbf{0},\omega^{-1}\boldsymbol{\sigma})$ denotes a zero-centered normal distribution, $\omega \in \mathbb{R}^+$ is a precision and $\boldsymbol{\sigma} \in \mathbb{R}^{3\times 3}$ is a covariance matrix. The precision ω is treated as a latent variable drawn from the Gamma distribution and it can be interpreted as an observation weight. Therefore the variables $\omega_{1:J}$ characterize the landmarks $\boldsymbol{X}_{1:J}$: the higher the better. Unfortunately, direct minimization of (6) using (8) is intractable. Therefore one has to adopt a ECM formalism: the negative log-likelihood is replaced with the expected complete-data negative log-likelihood conditioned by the observed data, $E_\omega[-\log P(\omega_{1:J},\boldsymbol{X}|\boldsymbol{X}]$. ECM alternates between an E-step and an several conditional M-steps, i.e. Algorithm 1.

Algorithm 1: Robust face frontalization (RFF).

Data: 3D landmark coordinates $\boldsymbol{X}_{1:J}$, 3D shape reconstruction matrix $\mathbf{U}$ and mean shape $\overline{\boldsymbol{M}}$.
Initialization: $\boldsymbol{s}=\mathbf{0}$, $\boldsymbol{\sigma}=\mathbf{I}$, $\overline{\omega}_{1:J}=1_{1:J}$ and $\mu=1$;
Compute $\boldsymbol{X}'_{1:J}$ and $\boldsymbol{V}'_{1:J}$ with (18), (19);
Compute ρ with (13);
Use $\boldsymbol{\sigma}=\mathbf{I}$ in (14) to estimate $\mathbf{R}$ in closed form;
Compute $\boldsymbol{T}$, $\boldsymbol{\sigma}$ and $\boldsymbol{s}$ with (15), (16), (17);
This yields $\boldsymbol{\theta}=(\rho,\mathbf{R},\boldsymbol{T},\boldsymbol{s},\boldsymbol{\sigma})$.
while $\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\| > \epsilon$ **do**
  E-step: Evaluate a, $b_{1:J}$ and $\overline{\omega}_{1:J}$ with (9), (10);
  Update $\boldsymbol{X}'_{1:J}$ and $\boldsymbol{V}'_{1:J}, \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{V}}$ with (18), (19). ;
  M-rigid-step: Evaluate the new rigid parameters and covariance $\rho^\star, \mathbf{R}^\star, \boldsymbol{T}^\star, \boldsymbol{\sigma}^\star$ with (12)-(16);
  M-non-rigid-step: Evaluate the new non-rigid parameters $\boldsymbol{s}^\star$ with (17);
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^\star$;
end
Result:
Optimal model parameters $\boldsymbol{\theta}^\star=(\rho^\star,\mathbf{R}^\star,\boldsymbol{T}^\star,\boldsymbol{s}^\star,\boldsymbol{\sigma}^\star)$
Frontalized landmarks $\boldsymbol{Y}_{1:J}=\rho^\star\mathbf{R}^\star\boldsymbol{X}_{1:J}+\boldsymbol{T}^\star$.

The E-step computes the parameters of the weights' posterior distributions $\mathcal{G}(\omega_j;a,b_j)$, namely

$$a=\mu+\frac{3}{2}, \quad b_j=1+\frac{\|\boldsymbol{E}_j\|^2_{\boldsymbol{\sigma}}}{2}, \tag{9}$$

from which the posterior means are evaluated:

$$\overline{\omega}_j=\mathrm{E}[\omega_j|\boldsymbol{E}_j]=\frac{a}{b_j}. \tag{10}$$

The M-step consists of the estimation of the model parameters, namely $\boldsymbol{\theta} = (\rho, \mathbf{R}, \boldsymbol{T}, \boldsymbol{\sigma}, \boldsymbol{s})$, via the minimization of:

$$Q(\rho, \mathbf{R}, \boldsymbol{T}, \boldsymbol{s}, \boldsymbol{\sigma}) = \sum_{j=1}^{J} \overline{\omega}_j \|\rho \mathbf{R} \boldsymbol{X}_j + \boldsymbol{T} - \mathbf{W}_j \boldsymbol{S}\|_{\boldsymbol{\sigma}}^2 + \log|\boldsymbol{\sigma}| + \kappa \boldsymbol{s}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{s}, \quad (11)$$

and the computation of the parameter μ of the Gamma distribution, where $\Psi(a) \approx \log a - 1/2a$ is the digamma function:

$$\mu = \Psi^{-1}\left(\Psi(a) - \frac{1}{n}\sum_{j=1}^{J} \log b_j\right). \quad (12)$$

The last term of (11) is a regularizer that forces $\boldsymbol{s}$ to correspond to a *valid* shape, i.e, the constraint of (2). The minimization of (11) with respect to the parameters yields the following conditional expressions:

$$\rho^\star = \left(\frac{\sum_{j=1}^{J} \overline{\omega}_j \boldsymbol{V}_j'^\top \boldsymbol{\sigma}^{-1} \boldsymbol{V}_j'}{\sum_{j=1}^{J} \overline{\omega}_j (\mathbf{R}\boldsymbol{X}_j')^\top \boldsymbol{\sigma}^{-1} (\mathbf{R}\boldsymbol{X}_j')}\right)^{\frac{1}{2}}, \quad (13)$$

$$\mathbf{R}^\star = \underset{\mathbf{R}}{\operatorname{argmin}} \sum_{j=1}^{J} (\overline{\omega}_j \|\boldsymbol{V}_j' - \rho^\star \mathbf{R} \boldsymbol{X}_j'\|_{\boldsymbol{\sigma}}^2), \quad (14)$$

$$\boldsymbol{T}^\star = \widetilde{\boldsymbol{V}} - \rho^\star \mathbf{R}^\star \widetilde{\boldsymbol{X}}, \quad (15)$$

$$\boldsymbol{\sigma}^\star = \frac{1}{J}\sum_{j=1}^{J} \overline{\omega}_j (\boldsymbol{V}_j' - \rho^\star \mathbf{R}^\star \boldsymbol{X}_j')(\boldsymbol{V}_j' - \rho^\star \mathbf{R}^\star \boldsymbol{X}_j')^\top, \quad (16)$$

$$\boldsymbol{s}^\star = \left(\sum_{j=1}^{J} \overline{\omega}_j \mathbf{U}_j^\top \boldsymbol{\sigma}^{\star -1} \mathbf{U} + \kappa \boldsymbol{\Lambda}^{-1}\right)^{-1} \left(\sum_{j=1}^{J} \overline{\omega}_j \mathbf{U}_j^\top \boldsymbol{\sigma}^{\star -1} (\rho^\star \mathbf{R}^\star \boldsymbol{X}_j + \boldsymbol{T}^\star - \overline{\boldsymbol{M}}_j)\right), \quad (17)$$

where $\boldsymbol{X}_j', \boldsymbol{V}_j', \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{V}}$ are computed with:

$$\boldsymbol{X}_j' = \boldsymbol{X}_j - \widetilde{\boldsymbol{X}}, \quad \widetilde{\boldsymbol{X}} = \frac{\sum_{j=1}^{J} \overline{\omega}_j \boldsymbol{X}_j}{\sum_{j=1}^{N} \overline{\omega}_j}, \quad (18)$$

$$\boldsymbol{V}_j' = \boldsymbol{V}_j - \widetilde{\boldsymbol{V}}, \quad \widetilde{\boldsymbol{V}} = \frac{\sum_{j=1}^{J} \overline{\omega}_j \boldsymbol{V}_j}{\sum_{j=1}^{N} \overline{\omega}_j} \quad (19)$$

$$\boldsymbol{V}_j = \mathbf{U}_j \boldsymbol{s}^\star + \overline{\boldsymbol{M}}_j. \quad (20)$$

The ECM procedure is summarized in Algorithm 1.

3.3 Dynamic face deformation inference

We now describe a dynamic model for estimating a time-varying deformable face. Let $\boldsymbol{Y}_{1:t}$ ($1:t$ is a shorthand for $1, 2, \ldots, t$) be the sequence of frontalized landmarks obtained with Algorithm 1, where $\boldsymbol{Y}_t \in \mathbb{R}^{3J}$ is the vector of frontalized landmarks at t. For the sake of clarity, we regroup (3), (4) and (5):

$$\boldsymbol{V}_t = \mathbf{W}\boldsymbol{S}_t + \boldsymbol{F}_t, \quad (21)$$

$$\boldsymbol{Y}_t = \boldsymbol{V}_t + \boldsymbol{D}_t, \quad (22)$$

$$\boldsymbol{Y}_{tj} = \rho_t \mathbf{R}_t \boldsymbol{X}_{tj} + \boldsymbol{T}_t, \; j \in \{1 \ldots J\}, \quad (23)$$

We assume that the sequences $\boldsymbol{S}_{1:t}$ and $\boldsymbol{V}_{1:t}$ are Markovian stochastic processes, each one with its own dynamic regime and interconnected via (21). Moreover, $\boldsymbol{V}_{1:t}$ and $\boldsymbol{Y}_{1:t}$ are interconnected via (22). The graphical model shown on Figure 3 describes the proposed *doubly-latent* LDS (DL-LDS). Probabilistically, this system can be described with the following conditional distributions:

$$p(\boldsymbol{S}_t|\boldsymbol{S}_{t-1}) = \mathcal{N}(\boldsymbol{S}_t; \boldsymbol{S}_{t-1}, \boldsymbol{\Gamma}_S), \quad (24)$$

$$p(\boldsymbol{V}_t|\boldsymbol{V}_{t-1}, \boldsymbol{S}_t) = \mathcal{N}(\boldsymbol{V}_t; \alpha \boldsymbol{V}_{t-1} + (1-\alpha)\mathbf{W}\boldsymbol{S}_t, \boldsymbol{\Gamma}_V), \quad (25)$$

$$p(\boldsymbol{Y}_t|\boldsymbol{V}_t) = \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{V}_t, \boldsymbol{\Sigma}_t), \quad (26)$$

where $\boldsymbol{\Gamma}_S \in \mathbb{R}^{(K+1)\times(K+1)}$, $\boldsymbol{\Gamma}_V \in \mathbb{R}^{3J\times 3J}$, $\boldsymbol{\Sigma}_t \in \mathbb{R}^{3J\times 3J}$ are covariance matrices, and where $\boldsymbol{\Sigma}_t = \mathbf{I}_{J\times J} \otimes \boldsymbol{\sigma}_t$ is obtained from (16) ($\otimes$ denotes the Kronecker product). The main difference between standard LDSs and the proposed DL-LDS resides in the fact that there are two interconnected latent variables, having their own dynamical regimes. Consequently, the transition probability of $\boldsymbol{V}_t$, (25) is conditioned both by $\boldsymbol{V}_{t-1}$ and by $\boldsymbol{S}_t$. The scalar $\alpha \in [0, 1]$ weights the relative importance of the vertex dynamics and of the vertex reconstruction from the current shape embedding.

We now show that the above DL-LDS can be cast into a standard LDS, i.e. the Kalman filter. Let the latent variable $\boldsymbol{Z} = [\boldsymbol{S}; \boldsymbol{V}] \in \mathbb{R}^{K+1+3J}$ be the concatenation of the two latent variables. In the particular case of our graphical model, we have:

$$p(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1}) = p(\boldsymbol{S}_t, \boldsymbol{V}_t|\boldsymbol{S}_{t-1}, \boldsymbol{V}_{t-1}) = p(\boldsymbol{V}_t|\boldsymbol{S}_t, \boldsymbol{V}_{t-1})\, p(\boldsymbol{S}_t|\boldsymbol{S}_{t-1}), \quad (27)$$

where the pdfs on the second row are given by (24) and (25). Let $p(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1}) = \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_t, \boldsymbol{\Gamma})$. By taking the logarithm of both sides of (27) and by identifying the

quadratic and linear terms, we obtain:

$$\boldsymbol{\mu}_t = \boldsymbol{\Gamma}\mathbf{A}\boldsymbol{Z}_{t-1} \tag{28}$$

$$\boldsymbol{\Gamma}^{-1} = \begin{pmatrix} \boldsymbol{\Gamma}_S^{-1} + (1-\alpha)^2\mathbf{W}^\top\boldsymbol{\Gamma}_V^{-1}\mathbf{W} & -(1-\alpha)\mathbf{W}^\top\boldsymbol{\Gamma}_V^{-1} \\ -(1-\alpha)\boldsymbol{\Gamma}_V^{-1}\mathbf{W} & \boldsymbol{\Gamma}_V^{-1} \end{pmatrix} \tag{29}$$

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{\Gamma}_S^{-1} & -\alpha(1-\alpha)\mathbf{W}^\top\boldsymbol{\Gamma}_V^{-1} \\ 0 & \alpha\boldsymbol{\Gamma}_V^{-1} \end{pmatrix} \tag{30}$$

To summarize, (24), (25) and (26) can be rewritten as:

$$p(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1}) = \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\Gamma}\mathbf{A}\boldsymbol{Z}_{t-1}, \boldsymbol{\Gamma}) \tag{31}$$

$$p(\boldsymbol{Y}_t|\boldsymbol{Z}_t) = \mathcal{N}(\boldsymbol{Y}_t; \mathbf{C}\boldsymbol{Z}_t, \boldsymbol{\Sigma}_t), \tag{32}$$

where matrix $\mathbf{C} \in \mathbb{R}^{3J\times(K+1+3J)}$ projects the concatenated latent-variable space onto the space of observed variables. Similarly, matrix $\overline{\mathbf{C}} \in \mathbb{R}^{K\times(K+1+3J)}$ projects the concatenated latent-variable space onto the space of the shape embedding. These two matrices write:

$$\mathbf{C} = \begin{pmatrix} \mathbf{0}_{3J\times(K+1)} & \mathbf{I}_{3J\times 3J} \end{pmatrix} \tag{33}$$

$$\overline{\mathbf{C}} = \begin{pmatrix} \mathbf{I}_{K\times K} & \mathbf{0}_{(3J+1)\times K} \end{pmatrix} \tag{34}$$

We now follow the standard Bayesian derivation of the Kalman filter. For this purpose, we need to evaluate the following posterior and prior distributions:

$$p(\boldsymbol{Z}_t|\boldsymbol{Y}_{1:t}) = \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\nu}_t, \boldsymbol{\Psi}_t), \tag{35}$$

$$p(\boldsymbol{Z}_1) = \mathcal{N}(\boldsymbol{Z}_1; \boldsymbol{\nu}_1, \boldsymbol{\Psi}_1), \tag{36}$$

where $\boldsymbol{\nu}_t \in \mathbb{R}^{K+1+3N}$ and $\boldsymbol{\Psi}_t \in \mathbb{R}^{(K+1+3N)\times(K+1+3N)}$ are the mean and covariance, respectively. Applying the standard derivation of an LDS we have:

$$p(\boldsymbol{Z}_t|\boldsymbol{Y}_{1:t})\, p(\boldsymbol{Y}_t|\boldsymbol{Y}_{1:t-1}) = p(\boldsymbol{Y}_t|\mathbf{C}\boldsymbol{Z}_t)\, p(\boldsymbol{Z}_t|\boldsymbol{Y}_{1:t-1}), \tag{37}$$

as well as the marginalization:

$$p(\boldsymbol{Z}_t|\boldsymbol{Y}_{1:t-1}) = \int p(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1}) p(\boldsymbol{Z}_{t-1}|\boldsymbol{Y}_{1:t-1}) d\boldsymbol{Z}_{t-1}. \tag{38}$$

The integral can then be evaluated making use of the results of Bishop (2006):

$$\int \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\Gamma}\mathbf{A}\boldsymbol{Z}_{t-1}, \boldsymbol{\Gamma}) \mathcal{N}(\boldsymbol{Z}_{t-1}; \boldsymbol{\nu}_{t-1}, \boldsymbol{\Psi}_{t-1}) d\boldsymbol{Z}_{t-1}$$

$$= \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\Gamma}\mathbf{A}\boldsymbol{\nu}_{t-1}, \mathbf{P}_{t-1}) \tag{39}$$

$$\text{with :} \quad \mathbf{P}_{t-1} = \boldsymbol{\Gamma}\mathbf{A}\boldsymbol{\Psi}_{t-1}\mathbf{A}^\top\boldsymbol{\Gamma}^\top + \boldsymbol{\Gamma} \tag{40}$$

We can now write (37) as:

$$\mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\nu}_t, \boldsymbol{\Psi}_t)\, p(\boldsymbol{Y}_t|\boldsymbol{Y}_{1:t-1}) \tag{41}$$

$$= \mathcal{N}(\boldsymbol{Y}_t; \mathbf{C}\boldsymbol{Z}_t, \boldsymbol{\Sigma}_t)\, \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\Gamma}\mathbf{A}\boldsymbol{\nu}_{t-1}, \mathbf{P}_{t-1}),$$

from which we obtain the following recursive formulas:

$$\boldsymbol{\nu}_t = \boldsymbol{\Gamma}\mathbf{A}\boldsymbol{\nu}_{t-1} + \mathbf{K}_t(\boldsymbol{Y}_t - \mathbf{C}\boldsymbol{\Gamma}\mathbf{A}\boldsymbol{\nu}_{t-1}) \tag{42}$$

$$\boldsymbol{\Psi}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{C})\mathbf{P}_{t-1} \tag{43}$$

$$\mathbf{K}_t = \mathbf{P}_{t-1}\mathbf{C}^\top(\mathbf{C}\mathbf{P}_{t-1}\mathbf{C}^\top + \boldsymbol{\Sigma}_t)^{-1} \tag{44}$$

$$p(\boldsymbol{Y}_t|\boldsymbol{Y}_{1:t-1}) = \mathcal{N}(\boldsymbol{Y}_t; \mathbf{C}\boldsymbol{\Gamma}\mathbf{A}\boldsymbol{\nu}_{t-1}, \mathbf{C}\mathbf{P}_{t-1}\mathbf{C}^\top + \boldsymbol{\Sigma}_t) \tag{45}$$

In order to initialize the above recursion, one needs to provide the mean and covariance of the prior distribution, $\boldsymbol{\nu}_1$ and $\boldsymbol{\Psi}_1$, as well as the covariances $\boldsymbol{\Gamma}_S$ and $\boldsymbol{\Gamma}_V$ associated with the dynamics of $\boldsymbol{S}$ and of $\boldsymbol{V}$, respectively, i.e. (24) and (25). Let $\boldsymbol{S}_1$ be the shape embedding at $t = 1$, which is provided by Algorithm 1. The face vector of vertex coordinates is therefore estimated with $\boldsymbol{V}_1 = \boldsymbol{W}\boldsymbol{S}_1$. We have:

$$\boldsymbol{\nu}_1 = \begin{pmatrix} \boldsymbol{S}_1 \\ \boldsymbol{V}_1 \end{pmatrix}, \; \boldsymbol{\Psi}_1 = \mathbf{I}, \; \mathbf{P}_1 = \mathbf{I}. \tag{46}$$

Algorithm 2 describes an implementation of the proposed doubly-latent LDS combined with the robust estimation of the rigid transformation required by face frontalization. The output of Algorithm 2 is a temporal sequence of estimated embedding and vertices:

$$\hat{\boldsymbol{s}}_t = \overline{\mathbf{C}}\boldsymbol{\nu}_t \quad t \in \{1 \dots T\}, \tag{47}$$

$$\hat{\boldsymbol{V}}_t = \mathbf{C}\boldsymbol{\nu}_t \quad t \in \{1 \dots T\}. \tag{48}$$

Algorithm 2: Dynamic face frontalization (DFF).

Data: Temporal sequence of input landmark coordinates $\boldsymbol{X}_{1:T} = (\boldsymbol{X}_1 \dots \boldsymbol{X}_T)$, with $\boldsymbol{X}_t = (\boldsymbol{X}_{t1} \dots \boldsymbol{X}_{tJ})$, 3D shape reconstruction matrix $\mathbf{U}$ and mean shape $\overline{\boldsymbol{M}}$, covariance matrices $\boldsymbol{\Gamma}_S$, $\boldsymbol{\Gamma}_V$ and scalar α.
Initialization: Use Algorithm 1 to Initialize the DL-LDS parameters $\boldsymbol{\nu}_1$ and $\boldsymbol{\Psi}_1$ with (46);
while $t = 2 \dots T$ **do**
  Rigid-pose: Use Algorithm 1 to compute $\boldsymbol{Y}_t$;
  Evaluate $\boldsymbol{\Sigma}_t = \mathbf{I}_{J\times J} \otimes \boldsymbol{\sigma}_t^\star$;
  DL-LDS-recursion: Apply the recursive formulas (40), (42), (43) and (44) to compute parameters $\boldsymbol{\nu}_t, \boldsymbol{\Psi}_t$ and the gain matrix $\mathbf{K}_t$;
  Evaluate $\boldsymbol{s}_t$, $\boldsymbol{V}_t$ with (47), (48);
end
Result: Temporal sequence of shape embedding $\hat{\boldsymbol{s}}_{1:T}$, shape vertices $\hat{\boldsymbol{V}}_{1:T}$, and covariances $\boldsymbol{\Psi}_{1:T}$.

3.4 Face warping

A frontal view of the face is computed in the following way. For convenience, the temporal index t is dropped. A frontal 3D shape is first computed with (3). The points $\boldsymbol{V}_{1:N}$ are the vertices of a 3D triangulated mesh and therefore the projection of this mesh onto the frontal image I_f form a 2D triangulated mesh; assuming orthographic projection, the image coordinates of a vertex $\boldsymbol{V}_n$ are (V_{n1}, V_{n2}). Let k_1, k_2 and k_3 be the vertex indexes of a mesh triangle. We now compute the barycentric coordinates, $(\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3$ of a pixel $(a_1, a_2) \in \mathbb{N}^2$ that lies inside that triangle, i.e. $0 \leq \beta_1, \beta_2, \beta_3 \leq 1$. These barycentric coordinates correspond to the solution of the following set of linear equations:

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \beta_1 \begin{pmatrix} V_{k_1 1} \\ V_{k_1 2} \end{pmatrix} + \beta_2 \begin{pmatrix} V_{k_2 1} \\ V_{k_2 2} \end{pmatrix} + \beta_3 \begin{pmatrix} V_{k_3 1} \\ V_{k_3 2} \end{pmatrix} \tag{49}$$

$$1 = \beta_1 + \beta_2 + \beta_3 \tag{50}$$

Once the barycentric coordinates are computed, the depth $A_3 \in \mathbb{R}$ associated with pixel $(a_1 \ a_2)^\top$ is computed by linear interpolation, namely:

$$A_3 = \beta_1 V_{k_1 3} + \beta_2 V_{k_2 3} + \beta_3 V_{k_3 3} \tag{51}$$

The above procedure is repeated for all the triangles and for all the points inside each triangle, thus obtaining a frontal dense depth map for each face pixel. Let $\boldsymbol{A} = (a_1 \ a_2 \ A_3)^\top$ be the current point of the frontal dense depth map thus obtained.

The final face frontalization step consists of *warping* the face's pixel colors from the input-image I_p onto a synthesized frontal image I_f. The rigid transformation that maps the 3D face, from a frontal centered coordinate frame back onto the input view, is the inverse of the pose, namely $\rho' = \rho^{-1}$, $\mathbf{R}' = \mathbf{R}^\top$, and $\boldsymbol{T}' = -\rho^{-1}\mathbf{R}^\top \boldsymbol{T}$. The dense depth map of the face can therefore be mapped back with

$$\begin{pmatrix} B_1 \\ B_2 \\ B_3 \end{pmatrix} = \rho' \mathbf{R}' \begin{pmatrix} a_1 \\ a_2 \\ A_3 \end{pmatrix} + \boldsymbol{T}' \tag{52}$$

Assuming scaled orthographic projection, the 2D pixel location $(b_1, b_2) \in I_p$ is computed from the real-valued coordinates $(b_1, b_2) = ([B_1], [B_2])$, where $[\cdot]$ is the *round* operator. Because of self occlusions and of quantization, (52) maps several points, $\boldsymbol{A}^{1:Q}$, at the same pixel location, but with different depth values $(b_1, b_2, B_3^{1:Q})$. Notice that only the depth-map point with the smallest depth value should visible in the input image. Consequently, vertices that are not visible in the input image don't have any photometric information associated with them and hence, they give rise to blank areas in the frontalized image. The final face frontalization step consists of synthesizing a frontal image:

$$I_f(a_1, a_2) = \begin{cases} I_p(b_1, b_2) & \text{if } B_3 = \min_q \{B_3^q\}_{q=1}^Q \\ \emptyset & \text{otherwise,} \end{cases} \tag{53}$$

where $\emptyset$ means that there is no photometric information available with that pixel. This is illustrated on Figure 4.

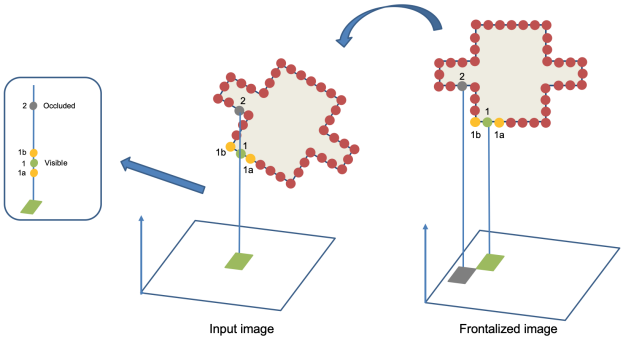


Fig. 4: When an object is rotated to appear frontal, some of its vertices have no associated photometric information in the synthesized frontal image, because they are not visible in the input image. In this example, vertices 1 and 2 are both visible in the frontalized image, but only 1 is visible in the input image. Because of quantization noise, the one-ring neighbors of 1, 1a and 1b, lie on the same line of sight as 1. We disregard this quantization effect and mark 1 as visible.

4 Implementation details

All the computations inside Algorithm 1 are in closed-form, with the notable exception of the estimation of the rotation matrix. The latter is parameterized with a unit quaternion Horn (1987), which allows one to reduce the number of rotation parameters, from nine to four, and to express the orthogonality constraints inside the rotation matrix in a much simpler way. The minimization (14) is carried out using a sequential least squares programming (SLSQP) solver[2] in combination with a root-finding software package Kraft (1988). The SLSQP minimizer found at the previous EM iteration is used to initialize the current EM iteration. At the start of EM, the closed-form method of Horn (1987) is used to initialize the rotation.

[2] https://docs.scipy.org/doc/scipy/reference/optimize.html

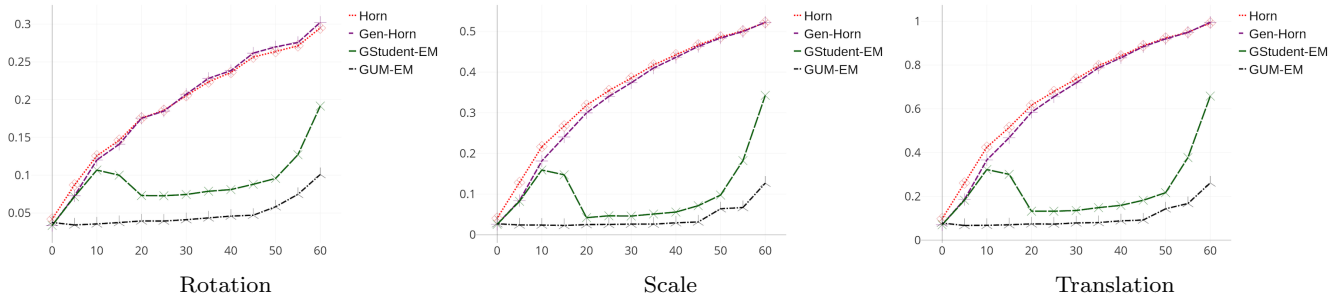


Fig. 5: The root mean-square error as a function of the percentage of outliers (i.e. landmark localization errors) averaged over 500 trials.

Algorithm 2 uses Algorithm 1 for intialization at $t = 1$. Then, at the following time steps, Algorithm 1 is used to compute the frontalized landmarks, which are then used to recursively estimate the parameters of the posterior (35) and the Kalman gain matrix (44). Note however, that the **M-non-rigid-step** of Algorithm 1 is not necessary because the shape embedding $\boldsymbol{s}$ is treated as a Gaussian variable. Instead, the **rigid-pose** step of Algorithm 2 uses the shape parameters evaluated at the previous time step. The value of α was empirically estimated and set to 0.06 in all the experiments. The covariance matrices $\boldsymbol{\Gamma}_S$ and $\boldsymbol{\Gamma}_V$ were experimentally evaluated from the results obtained with Algorithm 1 on a large dataset of faces.

In all the experiments we used the 3DFA method of Bulat and Tzimiropoulos (2016), as already mentioned in Section 2. 3DFA predicts $J = 68$ landmarks that may be prone to localization errors, i.e. outliers, in particular for non-frontal faces. We conducted a simulated experiment to show the effectiveness of the Student's t-distribution in the presence of outliers caused by 3DFA. For this purpose we considered a set of predicted 3D landmarks and we randomly simulated 500 rigid transformations. We added different noise types to the landmarks, as follows. For each trial we randomly split the landmarks into an inlier set and an outlier set. The inliers are corrupted by noise drawn from an anisotropic Gaussian distribution with a total variance $\lambda = 0.0025$ (the landmark coordinates are normalized to lie in the interval $[0, 1]$). The outlier noise is drawn from a uniform distribution whose volume is 1.5^3. We tested the following distributions: an isotropic Gaussian distribution (Horn), a full-covariance Gaussian distribution (Gen-Horn), a mixture distribution with a Gaussian component and a uniform component (GUM-EM), and the generalized Student's t-distribution used in the paper (GStudent-EM). Horn is named after the author of the well-known closed-form solution for estimating scale, rotation and translation between two 3D point sets Horn (1987). The plots of Figure 5 display the root mean-square error (RMSE) over 500 trials and for an increasing percentage of outliers (from 0% to 60%).

The proposed method also requires the parameters of an already trained deformable shape model, namely $\mathbf{U}, \overline{\boldsymbol{M}}$ in (2). For this purpose we combined two publicly available face models, Basel Shape Model (BSM) Paysan et al (2009) and Facewarehouse Cao et al (2014). BFM Paysan et al (2009) consists of a training set $\mathcal{M}^I = \{\boldsymbol{M}^I_m\}_{m=1}^{m=M^I}$ of $M^I = 200$ face scans of different identities. Each face in the dataset is frontally viewed and with a neutral expression. Each scan consists of a triangulated mesh composed of $N = 53490$ vertices. Both the vertices and the edges of the meshes are registered. Facewarehouse Cao et al (2014) consists of a training set $\mathcal{M}^E = \{\boldsymbol{M}^E_m\}_{m=1}^{m=M^E}$ of $M^E = 7050$ face scans that correspond to 150 identities and 47 expressions, with the same number of vertices N as BFM. The subjects were instructed to look frontally to the camera and to mimic 19 facial expressions as well as a neutral expression, from which 47 expressions were computed by linear blending. It should be noted that the 19 expressions correspond to emotions, e.g., mouth stretch, smile, anger, sadness, etc.

The identity and expression embeddings, $\boldsymbol{s}^I$ and $\boldsymbol{s}^E$, are of dimension $K^I = 199$ and $K^E = 29$, respectively. Therefore, a face mesh $\boldsymbol{V}$ is reconstructed from a linear combination of identity and expression:

$$\boldsymbol{V} = \mathbf{U}^I \boldsymbol{s}^I + \overline{\boldsymbol{M}}^I + \mathbf{U}^E \boldsymbol{s}^E + \overline{\boldsymbol{M}}^E \tag{54}$$

The above formula can be plugged into (11) whose minimization over $\boldsymbol{s}^I$ and $\boldsymbol{s}^E$ allows one to estimate the identity and expression embeddings of $\boldsymbol{V}$.

In this paper we are interested in processing a face sequence. Since the identity remains unchanged during

a sequence, the deformable face model just described is particularly interesting. Indeed, the identity embedding is estimating only once, at $t = 1$, which yields the following formulas to be used for the subsequent faces, i.e. for $t = 2 \dots T$:

$$\mathbf{U} = \mathbf{U}^E \tag{55}$$

$$\overline{\boldsymbol{M}} = \mathbf{U}^I \boldsymbol{s}_1^I + \overline{\boldsymbol{M}}^I + \overline{\boldsymbol{M}}^E \tag{56}$$

The processing time for a 256×256 face image is of 1.11 seconds on an Intel(R), Xeon(R) W-2145, 3.70GHz CPU equipped with a Quadro RTX 4000 GPU. This time decomposes as follows: 3D landmark extraction (0.48 s), pose estimation (0.02 s), model fitting (0.23 s), depth map interpolation and face warping (0.38 s).

5 Face frontalization benchmark

We now evaluate and benchmark the proposed face frontalization formulation based on a score that measures the correlation between a frontalized face and a ground-truth frontal image of the same face. For this purpose we use a dataset that contains pairs of frontal and profile videos of speaking participants for a large number of subjects. The evaluation consists of computing a metric between an image obtained by face frontalization of a profile view of a speaker, with an image containing a frontally-viewed face of the same speaker. It is important that the profile and frontal images are recorded with synchronized cameras in order to capture the same facial expression. Consequently, the proposed evaluation is based on image-to-image comparison. Several metrics were developed in the past for comparing two images, e.g. feature-based and pixel-based metrics. In this work we use the ZNCC score between two image regions, a measure that has successfully been used for stereo matching, e.g. Sun (2002). ZNCC is invariant to differences in brightness and contrast between the two images, due to the normalization with respect to mean and standard deviation.

Let $R_f(h, v) \subset I_f$ be a region of size $H \times V$ whose center coincides with pixel location (h, v) of a frontalized image I_f. Similarly, let $R_t(h, v) \subset I_t$ be a region of the same size and whose center coincides with pixel location (h, v) of a ground-truth image I_t. The ZNCC score between these two regions writes:

$$\text{ZNCC}(h, v, \delta h', \delta v') = \max_{\delta h, \delta v} \left\{ \frac{\text{Cov}\left[R_f(h, v), R_t(h + \delta h, v + \delta v)\right]}{\text{Var}\left[R_f(h, v)\right]^{1/2} \text{Var}\left[R_t(h + \delta h, v + \delta v)\right]^{1/2}} \right\}, \tag{57}$$

where $\text{Cov}[\cdot, \cdot]$ is the centered covariance between the two regions, $\text{Var}[\cdot]$ is the centered variance of a region, δh and δv are horizontal and vertical shifts, and $\delta h'$ and $\delta v'$ are the horizontal and vertical shifts that maximize the ZNCC score. ZNCC lies in the interval $[0, 1]$.

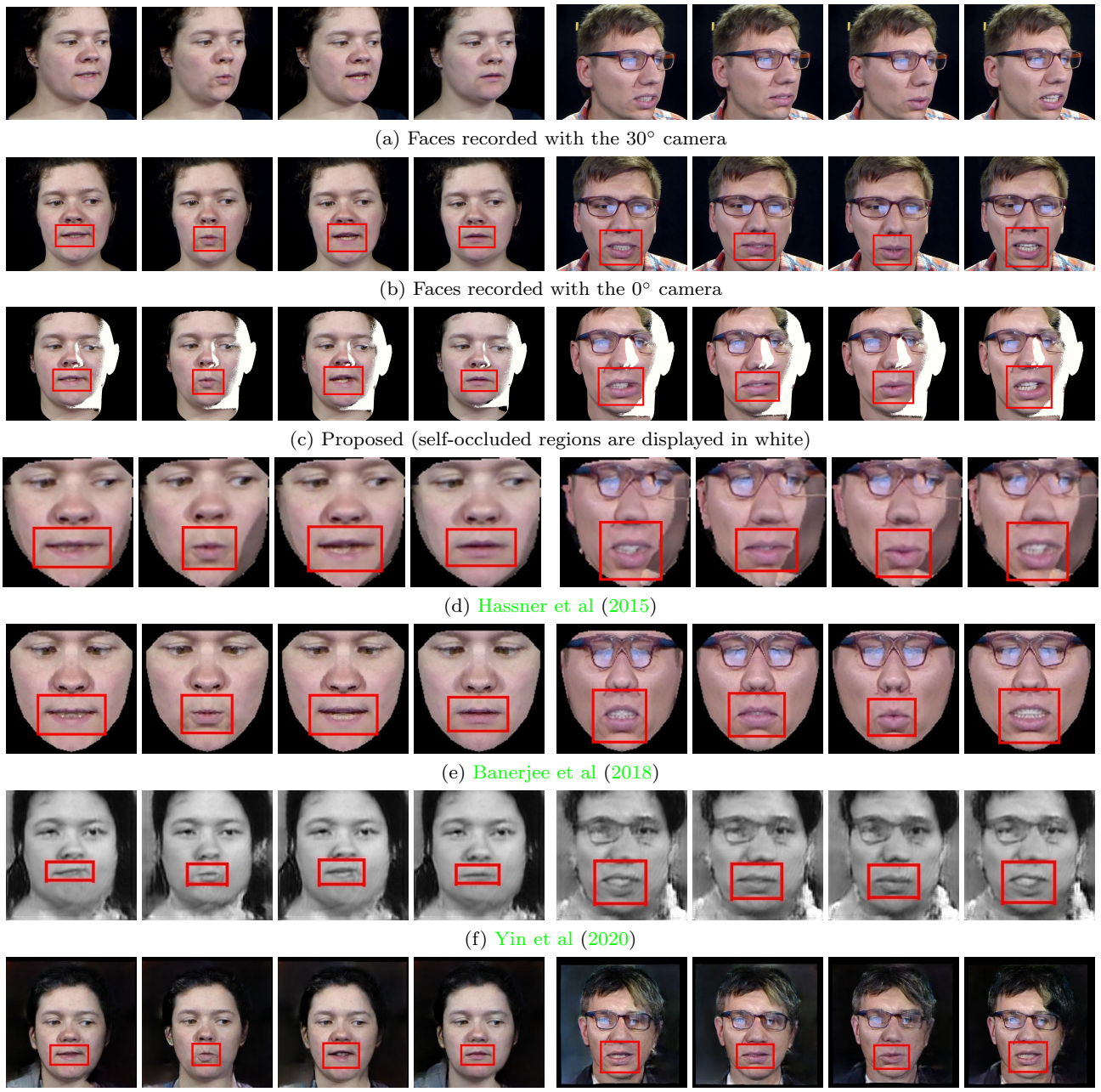
In order to evaluate the performance of the proposed frontalization method and to compare it with state-of-the-art methods, we used a publicly available dataset, namely the OuluVS2 dataset Anina et al (2015). This dataset targets the understanding of speech perception, more precisely, the analysis of non-rigid lip motions that are associated with speech production. The dataset was recorded in an office with ordinary (artificial and natural) lighting conditions. The recording setup consists of five synchronized cameras (2 MP, 30 FPS) placed at different points of view and with different orientations: 0°, 30°, 45°, 60°, 90°.

The dataset contains 5×106 videos recorded with 53 participants. Each participant was instructed to read loudly several text sequences displayed on a computer monitor placed slightly to the left and behind the 0° (frontal) camera. The displayed text consists of digit sequences, e.g. "one, seven, three, zero, two, nine", of phrases, e.g. "thank you", "have a good time", and "you are welcome", as well as of sequences from the TIMIT dataset, e.g. "agricultural products are unevenly distributed". While participants were asked to keep their heads still, natural uncontrolled head movements and body position changes were inevitable. As a consequence the actual head pose varies from one participant to another and there is no exact match between the head and camera orientations.

In practice, we evaluated the performance of the proposed method and we compared it with four state-of-the-art methods for which the code is publicly available, Hassner et al (2015); Banerjee et al (2018); Zhou et al (2020); Yin et al (2020). We applied the frontalization to images extracted from the videos recorded with the 30° camera (I_p) and compared the results with the "ground-truth", namely the corresponding images ex-

Method	Principle	ZNCC
Hassner et al	2D-to-3D fitting + symmetry	0.771
Banerjee et al	2D-to-3D fitting + symmetry	0.749
Zhou et al	2D-to-3D fitting + GAN	0.793
Yin et al	2D-to-2D mapping using GAN	0.769
Ma et al	2D-to-2D affine fitting	0.760
Kang et al	3D-to-3D robust fitting	0.831
Proposed	3D-to-3D robust/dynamic inference	**0.839**

Table 1: Mean ZNCC scores for all 53 participants of the OuluVS2 dataset. ZNCC lies in the interval $[0, 1]$.



(a) Faces recorded with the 30° camera

(b) Faces recorded with the 0° camera

(c) Proposed (self-occluded regions are displayed in white)

(d) Hassner et al (2015)

(e) Banerjee et al (2018)

(f) Yin et al (2020)

(g) Zhou et al (2020)

Fig. 6: Frontalization examples for participants #02 (left) and #21 (right) from the OuluVS2 dataset. The ZNCC scores correspond to the mouth bounding boxes shown in red. The estimated horizontal head orientation (yaw angle) is 24.9° and 40.6° for participant #2 and #21, respectively.

tracted from the videos recorded with the 0° camera (I_t). Notice that videos recorded with higher viewing angles, i.e. 45°, 60° and 90°, can be hardly exploited by a frontalization algorithm because half of the face is occluded. For each frontalized image I_f we extract the mouth region R_f and we search in the associated ground-truth image I_t for the best-matching region R_t. This provides a ZNCC score (57) for each query image I_p. Notice that (57) only cares about the horizontal and vertical shifts in the image plane and assumes that the frontalized face and the corresponding ground-truth frontal face share the same scale. In practice, different frontalization algorithms output faces at different scales. For this reason and for the sake of fairness, prior to applying (57), we extract facial landmarks from both the frontalized and ground-truth faces and we use a sub-

Part.	Yaw	Hassner et al	Banerjee et al	Zhou et al	Yin et al	Kang et al	Proposed
#31	19.1	0.905	0.856	0.822	0.875	**0.927**	***0.925***
#01	23.5	0.915	0.893	0.884	**0.921**	0.909	***0.918***
#02	24.9	0.888	0.878	0.929	0.881	**0.956**	***0.952***
#10	29.0	0.805	0.812	**0.873**	0.792	0.812	***0.829***
#23	30.0	0.810	**0.857**	0.819	0.817	***0.847***	0.843
#27	32.9	0.685	**0.852**	***0.824***	0.772	0.787	0.805
#19	37.8	***0.752***	0.650	0.662	0.677	**0.755**	**0.755**
#12	38.5	0.731	0.713	0.755	0.683	**0.770**	***0.766***
#21	40.6	0.632	0.743	0.653	0.673	**0.766**	***0.751***
Mean		0.791	0.801	0.802	0.787	***0.836***	**0.838**

Table 2: ZNCC scores for nine participants as a function of estimated yaw angle (in degrees) that corresponds to the horizontal head orientation computed with the proposed 3D head-pose estimator. For each participant, the best scores are in **bold** and the second best are in ***slanted bold***.

set of this set of landmarks to estimate the scale factor between the two faces. We do this for all the frontalization methods used in the comparison.

We used the 106 video pairs recorded with the 30° and 0° cameras, respectively, associated with the 53 participants of the OuluVS2 dataset. Each video contains 160 images, hence there are $106 \times 160 = 16,900$ image pairs in our benchmark. The mean ZNCC scores obtained with four methods, with Kang et al (2021), and with the proposed extension are shown in Table 1. We noticed that there were important discrepancies in method performance across participants. In order to better understand this phenomenon, we computed the mean ZNCC scores for nine participants and displayed these means as a function of the yaw angle, i.e. horizontal head orientation estimated with the proposed method, Table 2. One may notice that there is a wide range of yaw angles, from 19° to 40°, and that the performance gracefully decreases as the yaw angle increases. The proposed method yields results that are more consistent than the other methods, as the yaw angle increases.

The best performing methods are Kang et al (2021) and its dynamic extension. One remarks that the improvement of the dynamic model over Kang et al (2021) is minor, and this for the following reason. The dynamic FF uses 68 observed landmarks in order to update the deformable model. However the latter is composed of thousands of vertices: consequently, the vast majority of these vertices are not observed. This means that the innovation term in (42) affects a handful of the shape's vertices.

Examples of face frontalization obtained with our method and with four other methods, Hassner et al (2015); Banerjee et al (2018); Zhou et al (2020); Yin et al (2020), are shown on Figure 6: (a) input mages recorded with the 30° camera, (b) ground-truth images recorded with the 0° camera, (c)-(g) frontalization results. The ZNCC correlation scores correspond to the mouth region, shown in red. As already mentioned, both Hassner et al (2015) and Banerjee et al (2018) enforce facial symmetry as a post-processing frontalization step to compensate for the gaps caused by self occlusions. It is interesting to note that the more recent GAN-based methods, Zhou et al (2020); Yin et al (2020), yield results comparable with the traditional computer vision methods.

6 Lip reading benchmark

We also evaluated the ability of our method to improve the performance of lip reading and we compared it with other methods. For this purpose, we considered an isolated word recognition (IWR) task. The LRW (lip reading in the wild) dataset Chung and Zisserman (2016) consists of 500,000 videos of 500 English words uttered by 1,000 different speakers. Each video contains 29 frames and each target word is surrounded by context words. There are large inter-speaker variations in terms of head motions. To date, the best performing method for this 500-IWR task is based on the temporal convolutional network (TCN) model of Martinez et al (2020); Ma et al (2021a,b) which achieves a word classification score (WCS) of 87%. This lip-reading model and its variants use their own FF method which estimates a 3D affine mapping between the input face and a generic face model, Martinez et al (2020). Their FF is used as a preprocessing stage for training, validation and test. The authors don't provide a detailed description of the frontalization method that they use.

We performed the following 500-IWR experiments. In the first experiment we used the lip-reading model provided by the publicly available software packages

Training \ Testing	Hassner et al	Zhou et al	Yin et al	Ma et al	Proposed
Training with Ma et al (2021a)	60	59	20	87	82
Fine tuning with Zhou et al (2020)	60	72	20	84	80
Fine tuning with proposed	64	66	24	88	85

Table 3: Word classification scores (WCSs) in %. *First row:* The lip-reading model is trained with the built-in FF of Ma et al (2021a); *Second row:* the lip-reading model is fine tuned with the FF of Zhou et al (2020); *Third row:* The lip-reading model is fine tuned with the proposed FF method. For testing, we preprocessed the test images with the FF methods included in the comparison.

of Ma et al (2021a). This model is trained with their FF. In the second experiment we preprocessed a subset of the training dataset with the proposed dynamic FF method and we fine tuned the lip-reading model of Ma et al (2021a) on the 500-IWR task. For the purpose of fine tuning, for each one of the 500 words, we used 200 videos for training and 20 videos for validation, hence 100,000 training videos and 10,000 validation videos. Finally, we repeated the second experiment using Zhou et al (2020) for FF. In order to test these three models, we used the entire test dataset of LRW, namely 20 test videos for each one of the 500 words. The test videos were then preprocessed with each one of the FF models included in the benchmark: Table 3 shows the results obtained with 3×5 configurations corresponding to different train/test combinations. The proposed/Ma et al combination yields the best results: for this train/test combination, the WCS score is slightly increased, from 87 to 88, while the Zhou et al/Ma et al combination decreases the WCS score from 87 to 84.

The proposed frontalization model preserves facial expressions, hence the statistical properties of the training dataset are preserved. On the contrary, the GAN-based method of Zhou et al (2020) doesn't enjoy this Euclidean invariance. Consequently, the statistical distribution of the data used for fine tuning is modified. The model tends to overfit to the new distribution thus leading to a performance drop, as if the finely tuned model forgets what it was learned before. This phenomenon is referred to as catastrophic learning McCloskey and Cohen (1989), and is extensively investigated in continual learning.

7 Audio-visual speech enhancement

In this section we report experiments with using the proposed method in conjunction with AVSE. We start by summarizing the AVSE method based on a conditional variational auto-encoder (VAE) model Sadeghi et al (2020), which we denote AV-CVAE. The whole framework consists of two steps: training and testing (inference). At training, a prior distribution of clean speech is learned from the concatenation of a clean audio signal with an embedding of the associated lip images. At inference, clean speech is extracted from a noisy-speech signal and from a sequence of lip images: the learned prior distribution is combined with a noise model, whose parameters together with the parameters of the clean speech that were previously learned, are estimated following a variational expectation-maximization (VEM) procedure.

Given a dataset of complex-valued short-time Fourier transform (STFT) frames of a clean-speech signal, denoted $\boldsymbol{s}_t \in \mathbb{C}^F$, and the corresponding lip embedding obtained from a lip bounding-box cropped from the image of a speaker face, denoted $\boldsymbol{v}_t \in \mathbb{R}^M$, a latent-variable generative model is trained using the VAE framework. This involves defining a parametric distribution for the likelihood $p_\Theta(\boldsymbol{s}_t|\boldsymbol{z}_t, \boldsymbol{v}_t)$, and a parametric prior distribution for the latent code $\boldsymbol{z}_t \in \mathbb{R}^L$, $L \ll F$, $p_\Gamma(\boldsymbol{z}_t|\boldsymbol{v}_t)$. These distributions are implemented by some deep neural network architectures, whose parameters, $\{\Theta, \Gamma\}$, are learned following an amortized variational inference Kingma and Welling (2014), where an encoder network is introduced to approximate the intractable posterior distribution of the latent codes. Fig. 7 illustrates the AV-CVAE architecture. The main difference between this architecture and the one proposed in Sadeghi et al (2020); Sadeghi and Alameda-Pineda (2021) is the presence of a ResNet backbone from a pretrained model specialized for lip reading Martinez et al (2020).

With the parametric prior distribution for clean speech being learned, we consider an observation model as $\boldsymbol{o}_t = \boldsymbol{s}_t + \boldsymbol{b}_t$, in which $\boldsymbol{o}_t \in \mathbb{C}^F$ and $\boldsymbol{b}_t \in \mathbb{C}^F$ denote observed speech and noise, respectively. Considering an NMF-based model for noise, and combining it with the speech model, the set of NMF parameters are then learned by a variational inference procedure. Once learned, the clean speech estimate $\hat{\boldsymbol{s}}_t$ is obtained via a probabilistic Wiener filtering. More details can be found in Sadeghi et al (2020).

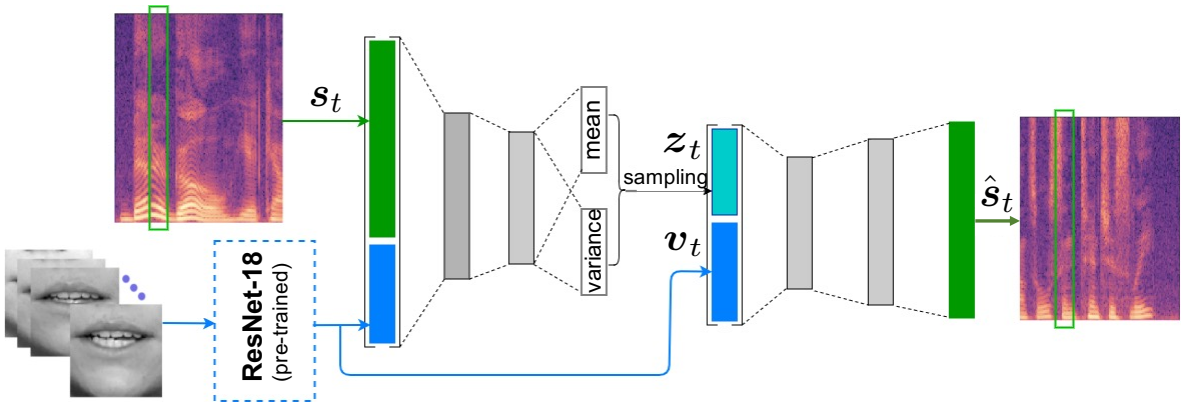


Fig. 7: AV-CVAE and ResNet-AV-CVAE architectures used in our speech enhancement experiments.

Measure	STOI $[0,1]\uparrow$					PESQ $[-0.5,4.5]\uparrow$					SI-SDR (dB) $\uparrow$				
SNR (dB)	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10
Noisy audio input	0.40	0.53	0.66	0.78	**0.86**	0.90	1.24	1.67	2.05	2.42	-15.92	-10.62	-5.44	-0.40	4.60
A-VAE Leglaive et al	0.41	0.56	0.70	**0.79**	0.85	0.93	1.51	2.02	2.43	2.73	-7.01	-0.29	5.08	9.41	12.74
AV-CVAE Sadeghi et al	0.42	0.57	0.69	**0.79**	0.84	1.02	1.56	2.06	2.42	2.73	-6.96	-0.04	5.01	9.06	12.25
Res-AV-CVAE-W/O-FF	0.41	0.55	0.67	0.77	0.83	1.02	1.53	1.99	2.35	2.70	-7.84	-0.60	4.68	8.81	12.30
Res-AV-CVAE-DA-ST-GAN Zhou et al	0.40	0.55	0.68	0.78	0.84	1.01	1.54	2.01	2.39	2.72	-7.92	-1.14	4.13	9.27	11.77
Res-AV-CVAE-DA-GAN Yin et al	0.39	0.55	0.66	0.68	0.72	0.76	1.42	1.87	1.66	1.96	-9.08	-0.45	3.88	4.55	5.23
Res-AV-CVAE-RFF Kang et al	**0.43**	0.58	0.71	**0.79**	0.85	1.12	1.69	2.13	**2.48**	**2.77**	**-6.30**	0.10	5.24	9.30	12.60
Res-AV-CVAE-DFF	**0.43**	**0.60**	**0.73**	**0.79**	0.85	**1.13**	**1.71**	**2.20**	**2.48**	**2.77**	-6.35	**0.28**	**5.87**	**9.42**	**12.77**

Table 4: Average STOI, PESQ, SI-SDR values.

All the experiments reported below use the MEAD dataset Wang et al (2020) which contains short videos of talking faces with large-scale facial expressions. For all 46 publicly available participants, there are recordings of eight different emotions at three different intensity levels and seven camera viewpoints. Many participants have natural head motions, which challenges state-of-the-art AVSE. Among all videos, we select the videos of all emotion categories taken at the frontal view and at the level 3 (the highest) of emotion intensity. These high-intensity emotions are associated with large head movements and exaggerated lip motions, thus allowing to assess the effect of head movements on the performance of speech enhancement. In total, there are around 5 hours of videos for training, 0.7 hours for validation and 0.7 hours for testing.

We process the input videos with four different FF methods in order to compare their effectiveness of removing head movements and hence of improving the quality of the speech output: the GAN based methods Zhou et al (2020) and Yin et al (2020), denoted ST-GAN and DA-GAN, respectively, the method of Kang et al (2021) that corresponds to Algorithm 1, denoted RFF, and the dynamic method that corresponds to Algorithm 2, denoted DFF. Additionally, we consider the case of directly using the raw input without any form of face frontalization, denoted W/O-FF. For all these cases we crop the lip region, yielding 67×67 images, which are then converted to gray scale and normalized to facilitate the downstream processing.

We consider three speech enhancement pipelines, all based on VAEs. The Audio-only VAE (A-VAE), Leglaive et al (2018) has an encoder and a decoder composed of fully-connected layers. The extracted audio feature vector is of size $F = 513$ whereas the latent space is of size $L = 32$. The AV-CVAE model Sadeghi et al (2020) shares a similar encoder-decoder architecture as A-VAE, with the additional fully-connected layers to encode the visual information. Furthermore, we propose to use a ResNet backbone specially trained for lip reading Martinez et al (2020) (shown in a dashed box in Figure 7) for visual feature extraction. The backbone follows the standard design of ResNet-18 He et al (2016) except for the first convolutional layer, which is replaced by a 3D convolutional layer to incorporate temporal information from neighbouring frames. This variant is denoted as Res-AV-CVAE. In practice, the dimension of the visual embedding is $M = 128$.

All the VAEs are trained in an end-to-end manner. A-VAE is trained on audio data. AV-CVAE Sadeghi et al (2020) is fine-tuned with the MEAD dataset, whereas Res-AV-CVAE is trained from scratch using MEAD.

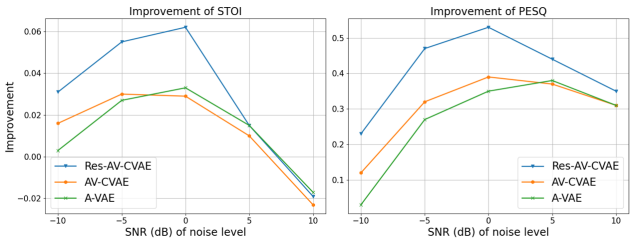


Fig. 8: Performance comparison of A-VAE, AV-CVAE and Res-AV-CVAE based on STOI (left) and PESQ (right).

Note that the ResNet backbone is frozen without requiring the gradients. Hence, it is a static feature extractor. We set $5e^{-5}$ as the learning rate for the finetuning model and $1e^{-4}$ for the training from scratch. The Adam optimizer was used with a batch size of 128. We also applied early stopping with a patience of 10 epochs. Note that we trained and tested one model with one specific lip preprocessing method at a time. At test time, noise from the DEMAND dataset Thiemann et al (2013) is combined with the clean speech to construct the audio input. There are five noise levels for each type of noise, namely -10 dB, -5 dB, 0 dB, 5 dB and 10 dB. Three standard speech enhancement metrics are used for quantitative evaluation: the *scale-invariant signal-to-distortion ratio* (SI-SDR) Le Roux et al (2019), the *short-time objective intelligibility* (STOI) Taal et al (2011) and the *perceptual evaluation of speech quality* (PESQ) Rix et al (2001). SI-SDR is measured in decibels (dB), while STOI and PESQ values are in the range $[0, 1]$ and $[-0.5, 4.5]$, respectively (the higher the better).

We start with evaluating the impact of different frontalization methods on AVSE performance, i.e. Table 4, where the average scores for different levels of noise (SNR) are presented. Selecting RFF and DFF – the best-performing methods – as examples, we remark that the difference between with and without frontalization is significant. This confirms that the head motions interfere the processing of visual speech patterns. In other words, separating the rigid head movements from the non-rigid lip deformations allows the model to learn a better clean speech model. Moreover, the comparison between A-VAE and Res-AV-CVAE with RFF/DFF further validates the contribution of the visual modality. In addition, DFF demonstrates a better performance than RFF, especially in terms of the speech intelligibility score STOI at high noise levels. However, one should note that the VAE models used in this paper do not incorporate the temporal dynamics of the audio and visual data, and rather process the data time frames independently. Using a dynamical VAE model would lead to even higher performance gain in DFF compared to RFF.

The choice of the face frontalization method is important. While RFF/DFF offers significant improvements, ST-GAN yields a minor difference compared to the case with head movements. Indeed, GAN-based image generation models have no theoretical guarantee for preserving the lip shape – they add some form of visual noise, which neutralizes the gain of frontalization. This explanation is also supported by the results of DA-GAN: its performance is falling far behind the other methods. As the results of ST-GAN are conditioned on the transformation-based process, the model possesses a prior knowledge about the frontalized face. Moreover, the direct mapping from an arbitrary viewpoint to a frontal view of DA-GAN introduces even more dramatic modifications in the lip shape. Thus, the model has more difficulties to learn the correct speech patterns from lip movements.

We then compare the performance of different VAE architectures in Figure 8, where the improvement of scores are shown as a function of different levels of noise (SNR). More precisely, the improvement refers to the difference between the score obtained by using the raw noisy speech and those obtained by using the enhanced speech. First, it is remarkable to see that Res-AV-CVAE significantly outperforms AV-CVAE, showing the gain of using a more powerful feature extractor. Second, we observe that with a noise level in the range $[-5, 0]$ dB, the Res-AV-CVAE model reaches an optimal stage (a peak in the curve) for fusing the audio-visual data. That is, with the noise level going higher (smaller SNR), the audio would be too corrupted to be enhanced and the visual contribution is significant. In contrast, with the noise level going lower (higher SNR), the importance of the visual data is decreasing and the already clean speech becomes harder to be enhanced. While the superior performance of the Res-AV-CVAE models is more significant at high noise levels, it is quite remarkable to observe that Res-AV-CVAE-RFF performs almost equally well as A-VAE for low noise levels. These experiments confirm the complementary roles of the visual and audio modalities for the task of speech enhancement.

To give an insight on the impact of removing head movements, Figure 2 shows the horizontal and vertical displacements of a landmark located on the upper lip. Both the vertical and horizontal trajectories of this lip landmark are strongly affected by head motions. In the light of this experiment, one may interpret the process of separating rigid head movements and non-rigid lip

movements, as a way of extracting clean visual-speech information from the raw videos.

8 Conclusions

Shape as defined by Kendall (1989), is the geometric information that remains once an object has been normalized with respect to rotation, scaling and translation. The proposed face frontalization methodology follows this definition and hence it guarantees that face geometric information, i.e. non-rigid facial deformations, is preserved. This stays in contrast with state-of-the-art DNN-based frontalization methods that learn millions of parameters without the theoretical guarantee that they faithfully preserve facial deformations.

We conducted several experiments in order to analyze the effect of frontalization onto visual speech processing, whose success critically relies on the analysis of non-rigid mouth motions, e.g. lip reading. For this purpose, we used three datasets, OuluVS2, LRW, and MEAD.

We proposed an evaluation pipeline that consists of measuring the ZNCC score between a frontalized face and a frontal view of the same face. We compared our method with four state-of-the-art methods that use various geometric and DNN models. This benchmark reveals that the proposed method performs better than the other methods in preserving the shape of the mouth.

The LRW and MEAD datasets contain videos of persons uttering speech. Unlike the OuluVS2 participants, who keep their heads in a fixed position and orientation, the LRW and MEAD participants perform head motions – a natural human behavior. We combined our frontalization method with two speech processing tasks, lip reading and speech enhancement, and we thoroughly analyzed its effect onto two scores: word classification and speech intelligibility. These experiments reveal that these scores are improved significantly with respect with both classical geometric models and GAN models.

It is interesting to remark that the proposed formulation may well be viewed as a method for separating rigid head motions from non-rigid facial expressions. This is useful, not only for improving the performance of visual speech, but for a number of other tasks that involve the analysis of facial expressions in realistic scenarios.

References

Abdelaziz AH (2017) NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition. In: INTERSPEECH, pp 3752–3756 2

Adeel A, Gogate M, Hussain A, Whitmer WM (2019) Lip-reading driven deep learning approach for speech enhancement. IEEE Transactions on Emerging Topics in Computational Intelligence 1

Adeel A, Gogate M, Hussain A, Whitmer WM (2021) Lip-reading driven deep learning approach for speech enhancement. IEEE Transactions on Emerging Topics in Computational Intelligence 5(3):481–490 2

Anina I, Zhou Z, Zhao G, Pietikäinen M (2015) OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis. In: International Conference on Automatic Face and Gesture Recognition, IEEE, vol 1, pp 1–5 2, 11

Banerjee S, Brogan J, Krizaj J, Bharati A, Webster BR, Struc V, Flynn PJ, Scheirer WJ (2018) To frontalize or not to frontalize: Do we really need elaborate pre-processing to improve face recognition? In: IEEE Winter Conference on Applications of Computer Vision, pp 20–29 1, 3, 4, 11, 12, 13

Baumberg A (1998) Hierarchical shape fitting using an iterated linear filter. Image and Vision Computing 16(5):329–335 4

Bishop C (2006) Pattern Recognition and Machine Learning. Springer 8

Blanz V, Vetter T (1999) A morphable model for the synthesis of 3D faces. In: ACM SIGGRAPH, vol 99, pp 187–194 3, 5

Bulat A, Tzimiropoulos G (2016) Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3DFAW) challenge. In: European Conference on Computer Vision Workshops, Springer, pp 616–624 4, 10

Bulat A, Tzimiropoulos G (2017) How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: IEEE International Conference on Computer Vision, pp 1021–1030 4

Cao C, Weng Y, Zhou S, Tong Y, Zhou K (2014) Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics 20(3):413–425 10

Cheng S, Ma P, Tzimiropoulos G, Petridis S, Bulat A, Shen J, Pantic M (2020) Towards pose-invariant lip-reading. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp 4357–4361 1

Chung JS, Zisserman A (2016) Lip reading in the wild. In: Asian Conference on Computer Vision, pp 87–103

2, 13
Deng J, Zhou Y, Cheng S, Zaferiou S (2018) Cascade multi-view hourglass model for robust 3d face alignment. In: IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, pp 399–403 4
Feng Y, Wu F, Shao X, Wang Y, Zhou X (2018) Joint 3D face reconstruction and dense alignment with position map regression network. In: European Conference on Computer Vision, pp 534–551 4
Fernandez-Lopez A, Sukno FM (2018) Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing 78:53–72 1, 2
Ferrari C, Lisanti G, Berretti S, Del Bimbo A (2016) Effective 3D based frontalization for unconstrained face recognition. In: International Conference on Pattern Recognition, IEEE, pp 1047–1052 4
Forbes F, Wraith D (2014) A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. Statistics and Computing 24(6):971–984 3
Gao W, Cao B, Shan S, Chen X, Zhou D, Zhang X, Zhao D (2007) The CAS-PEAL large-scale chinese face database and baseline evaluations. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 38(1):149–161 4
Gross R, Matthews I, Cohn J, Kanade T, Baker S (2010) Multi-PIE. Image and Vision Computing 28(5):807–813 4
Hassner T, Harel S, Paz E, Enbar R (2015) Effective face frontalization in unconstrained images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 4295–4304 3, 4, 11, 12, 13, 14
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778 15
Horn BK (1987) Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America A 4(4):629–642 9, 10
Hou JC, Wang SS, Lai YH, Tsao Y, Chang HW, Wang HM (2018) Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence 2(2):117–128 2
Huang GB, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: A database for studying face recognition in unconstrained environments 4
Huang R, Zhang S, Li T, He R (2017) Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In: IEEE International Conference on Computer Vision, pp 2439–2448 3
Jiang L, Wu XJ, Kittler J (2019) Dual attention mobdensenet (damdnet) for robust 3D face alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops 4
Kang Z, Horaud R, Sadeghi M (2021) Robust face frontalization for visual speech recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp 2485–2495 2, 5, 11, 13, 15
Kendall DG (1989) A survey of the statistical theory of shape. Statistical Science 4(2):87–99 17
Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: ICLR 14
Kraft D (1988) A software package for sequential quadratic programming. Tech. Rep. DFVLR-FB 88-28, DLR German Aerospace Center – Institute for Flight Mechanics, Koln, Germany 9
Le Roux J, Wisdom S, Erdogan H, Hershey JR (2019) SDR–half-baked or well done? In: ICASSP 16
Lee S, Kang J, Shin J, Paik J (2007) Hierarchical active shape model with motion prediction for real-time tracking of non-rigid objects. IET Computer Vision 1(1):17–24 4
Leglaive S, Girin L, Horaud R (2018) A variance modeling framework based on variational autoencoders for speech enhancement. In: MLSP 15
Ma P, Martinez B, Petridis S, Pantic M (2021a) Towards practical lipreading with distilled and efficient models. In: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp 7608–7612 2, 3, 11, 13, 14
Ma P, Wang Y, Shen J, Petridis S, Pantic M (2021b) Lip-reading with densely connected temporal convolutional networks. In: IEEE Winter Conference on Applications of Computer Vision 13
Martinez B, Ma P, Petridis S, Pantic M (2020) Lipreading using temporal convolutional networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp 6319–6323 1, 13, 14, 15
McClave EZ (2000) Linguistic functions of head movements in the context of speech. Journal of Pragmatics 32(7):855–878 2
McCloskey M, Cohen NJ (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol 24, Elsevier, pp 109–165 14
Michelsanti D, Tan ZH, Zhang SX, Xu Y, Yu M, Yu D, Jensen J (2021) An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29:1368–1396 2
Ning X, Duan P, Li W, Zhang S (2020) Real-time 3d face alignment using an encoder-decoder network

with an efficient deconvolution layer. IEEE Signal Processing Letters 27:1944–1948 4

Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T (2009) A 3D face model for pose and illumination invariant face recognition. In: IEEE International Conference on Advanced Video and Signal Based Surveillance, pp 296–301 10

Peel D, McLachlan GJ (2000) Robust mixture modelling using the t distribution. Statistics and Computing 10(4):339–348 3

Pei E, Oveneke MC, Zhao Y, Jiang D, Sahli H (2020) Monocular 3d facial expression features for continuous affect recognition. IEEE Transactions on Multimedia 1

Prabhu U, Seshadri K, Savvides M (2010) Automatic facial landmark tracking in video sequences using Kalman filter assisted active shape models. In: European Conference on Computer Vision, Springer, pp 86–99 4

Ravikumar N, Gooya A, Çimen S, Frangi AF, Taylor ZA (2018) Group-wise similarity registration of point sets using Student's t-mixture model for statistical shape models. Medical Image Analysis 44:156–176 3

Rix AW, Beerends JG, Hollier MP, Hekstra AP (2001) Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: ICASSP 16

Rong C, Zhang X, Lin Y (2020) Feature-improving generative adversarial network for face frontalization. IEEE Access 8:68,842–68,851 3

Sadeghi M, Alameda-Pineda X (2021) Mixture of inference networks for VAE-based audio-visual speech enhancement. IEEE Transactions on Signal Processing 69:1899–1909 3, 14

Sadeghi M, Leglaive S, Alameda-Pineda X, Girin L, Horaud R (2020) Audio-visual speech enhancement using conditional variational auto-encoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28:1788–1800 2, 3, 14, 15

Sariyanidi E, Zampella CJ, Schultz RT, Tunc B (2020) Can facial pose and expression be separated with weak perspective camera? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7173–7182 5

Schultz T, Wand M, Hueber T, Krusienski DJ, Herff C, Brumberg JS (2017) Biosignal-based spoken communication: A survey. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(12):2257–2271 2

Sun C (2002) Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques. International Journal of Computer Vision 47(1-3):99–117 3, 11

Taal CH, Hendriks RC, Heusdens R, Jensen J (2011) An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans Audio, Speech, Language Process 19(7):2125–2136 16

Thiemann J, Ito N, Vincent E (2013) Demand: a collection of multi-channel recordings of acoustic noise in diverse environments. In: Proc. Meetings Acoust, pp 1–6 16

Tran L, Yin X, Liu X (2017) Disentangled representation learning GAN for pose-invariant face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1415–1424 3

Tu X, Zhao J, Jiang Z, Luo Y, Xie M, Zhao Y, He L, Ma Z, Feng J (2020) 3d face reconstruction from a single image assisted by 2d face images in the wild. IEEE Transactions on Multimedia 23:1160–1172 4

Wang K, Wu Q, Song L, Yang Z, Wu W, Qian C, He R, Qiao Y, Loy CC (2020) MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In: ECCV, Springer 15

Yim J, Jung H, Yoo B, Choi C, Park D, Kim J (2015) Rotating your face using multi-task deep neural network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 676–684 1, 3

Yin X, Yu X, Sohn K, Liu X, Chandraker M (2017) Towards large-pose face frontalization in the wild. In: IEEE International Conference on Computer Vision, pp 3990–3999 3

Yin Y, Jiang S, Robinson JP, Fu Y (2020) Dual-attention GAN for large-pose face frontalization. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp 24–31 3, 4, 11, 12, 13, 14, 15

Zhang Z, Chen X, Wang B, Hu G, Zuo W, Hancock ER (2019) Face frontalization using an appearance-flow-based convolutional neural network. IEEE Transactions on Image Processing 28(5):2187–2199 3, 4

Zhang Z, Liang R, Chen X, Xu X, Hu G, Zuo W, Hancock ER (2021) Semi-supervised face frontalization in the wild. IEEE Transactions on Information Forensics and Security 16:909–922 3, 4

Zhao J, Cheng Y, Xu Y, Xiong L, Li J, Zhao F, Jayashree K, Pranata S, Shen S, Xing J, et al (2018) Towards pose invariant face recognition in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2207–2216 1, 3

Zhou E, Cao Z, Sun J (2018) Gridface: Face rectification via learning local homography transformations. In: Proceedings of the European Conference on Computer Vision 1, 4

Zhou H, Liu J, Liu Z, Liu Y, Wang X (2020) Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In: CVPR 1, 3, 4, 11, 12,

13, 14, 15

Zhou Z, Zheng J, Dai Y, Zhou Z, Chen S (2014) Robust non-rigid point set registration using student's-t mixture model. PloS one 9(3):e91,381 3

Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2223–2232 4

Zhu X, Lei Z, Yan J, Yi D, Li SZ (2015) High-fidelity pose and expression normalization for face recognition in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 787–796 1, 4

Zhu X, Lei Z, Liu X, Shi H, Li SZ (2016) Face alignment across large poses: a 3D solution. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 146–155 4

Zhu X, Liu X, Lei Z, Li SZ (2019) Face alignment in full pose range: A 3d total solution. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(1):78–92 4