SPECTRE: Visual Speech-Informed Perceptual 3D Facial Expression Reconstruction from Videos

Panagiotis P. Filntisis[1] George Retsinas [1,2] Foivos Paraperas-Papantoniou[5] Athanasios Katsamanis[6]
Anastasios Roussos[3,4] Petros Maragos[1,2]
[1]Institute of Robotics, Athena Research Center, 15125 Maroussi, Greece
[2]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[3]Institute of Computer Science (ICS), Foundation for Research & Technology - Hellas (FORTH), Greece
[4]College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK
[5]Imperial College London, UK
[6]Institute for Language and Speech Processing, Athena R.C., Greece

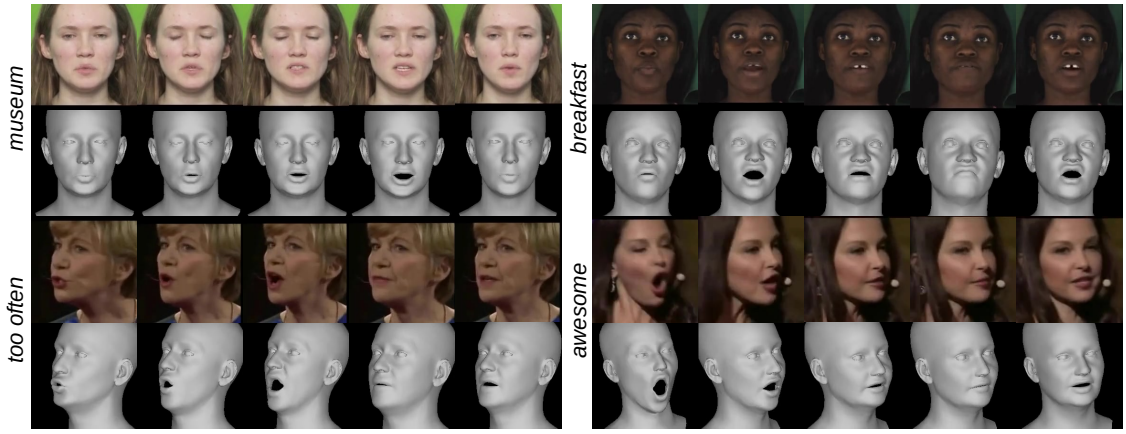

Figure 1. Our method SPECTRE performs visual-speech aware 3D reconstruction so that speech perception from the original footage is preserved in the reconstructed talking head. On the left we include the word/phrase being said for each example. Please zoom-in for details and refer to the Suppl. Material for videos of these results.

Abstract

The recent state of the art on monocular 3D face reconstruction from image data has made some impressive advancements, thanks to the advent of Deep Learning. However, it has mostly focused on input coming from a single RGB image, overlooking the following important factors: a) Nowadays, the vast majority of facial image data of interest do not originate from single images but rather from videos, which contain rich dynamic information. b) Furthermore, these videos typically capture individuals in some form of verbal communication (public talks, teleconferences, audiovisual human-computer interactions, interviews, monologues/dialogues in movies, etc). When existing 3D face reconstruction methods are applied in such videos, the artifacts in the reconstruction of the shape and motion of the mouth area are often severe, since they do not match well with the speech audio.

To overcome the aforementioned limitations, we present the first method for visual speech-informed perceptual reconstruction of 3D mouth expressions. We do this by proposing a "lipreading" loss, which guides the fitting process so that the elicited perception from the 3D reconstructed talking head resembles that of the original video footage. We demonstrate that, interestingly, the lipreading loss is better suited for 3D reconstruction of mouth movements compared to traditional landmark losses, and even direct 3D supervision. Furthermore, the devised method does not rely on any text transcriptions or corresponding audio, rendering it ideal for training in unlabeled datasets. We verify the efficiency of our method through objective evaluations on three large-scale datasets, as well as subjective evaluation with two web-based user studies. Project webpage: https://filby89.github.io/spectre/

1. Introduction

During the last years, Deep Learning frameworks have succeeded in significantly increasing the accuracy of monocular 3D face reconstruction, even in cases of unconstrained image data. The current state of the art is able

to robustly reconstruct fine details of the 3D facial geometry as well as yield a reliable estimation of the captured subject's facial anatomy. This is beneficial for multiple applications, such as augmented reality, performance capture, visual effects, photo-realistic video synthesis, human-computer interaction and personalized avatars, to name but a few. greedy the On the other hand, the vast majority of existing methods focus on 3D face reconstruction from a single RGB image, without exploiting the rich dynamic information that is inherent in humans' faces, especially during speech. But even the few methods that include some sort of dynamics modelling to reconstruct facial videos, do not explicitly model the strong correlation between mouth motions and articulated speech. At the same time, most facial videos of interest capture individuals involved in some form of verbal communication. When existing 3D face reconstruction methods are applied in this kind of videos, the artifacts in the reconstruction of the shape and motion of the mouth area are often severe and overwhelming in terms of human perception; the movements of the mouth that correspond to speech are not captured well.

Arguably, a crucial factor for the limitations of existing methods is the fact that most methods use weak 2D supervision from landmarks predicted by face alignment methods as a form of guidance [12, 25, 39, 52, 58, 61, 62, 69]. While these landmarks can yield a coarse estimation of the facial shape, they fail to provide an accurate representation of the expressive details of a highly-deformable mouth region. It is also important to note that the shapes of the human mouth are perceptually correlated with speech and the realism of a 3D talking head is tightly coupled with the uttered sentence. For example, a 3D model that talks without the lips closing when uttering the bi-labial consonants (i.e., /m/, /p/, and /b/), or with no lip-roundness when uttering a rounded-vowel (such as /o/ /u/) has a poor perceived naturalness. EMOCA [21] made some important advancements in terms of the expressivity of the 3D reconstructed head, however the perceptual emotional consistency loss only affected the movements that correspond to facial expressions. Furthermore, the estimation of jaw articulation was not included in the model, resulting often in poor reconstructions.

We conclude that, although speech perception from reconstructed 3D faces is important for various applications (e.g., augmented and virtual reality, gaming, affective avatars etc.) [37, 47, 57], it is a commonly overlooked parameter in the existing literature. To overcome the limitations of the existing literature, this work tackles the problem of monocular 3D face reconstruction from a video, with a strong focus on the mouth area and its expressions and movements that are connected with speech articulation. We highlight and address the fact that an accurate 3D reconstruction of a human talking in a video should retain those mouth expressions and movements that humans perceive to correspond to speech. Our method, dubbed *SPECTRE*, leverages a SoTA model of lip reading to minimize the "speech-informed perceptual" distance between the rendered and the original input video. Our main contributions can be summarized as follows:

• We design and implement the first (to our knowledge) method for perceptual 3D reconstruction of human faces focusing on speech **without the need for text transcriptions of the corresponding audio, or costly 3D annotations**.

• We propose a perceptual "lipreading" loss based on deep features, minimizing the perceptual distance of speech-related lip movements between the original and reconstructed (through a differentiable 3D face renderer) videos.

• We conduct experiments over the effectiveness of deep features against traditional geometric based metrics and showcase numerous examples where SPECTRE significantly outperforms other methods in speech-aligned mouth perceptibility, **as it can be clearly seen in the Suppl. video**. Our proposed system also generalizes well to other datasets, as demonstrated by our **cross-dataset experiments**.

• We make our source code and models publicly available at https://github.com/filby89/spectre.

2. Related Work

3D Models: There is extensive literature in the fields of computer vision and graphics for creating and reconstructing 3D face models from various input sources (RGB, Depth) [24, 73]. 3D Morphable Models are by far the most widely-used choice, since they offer compact representations as well as a convenient decoupling of expression and identity variation, allowing better manipulation. The traditional 3DMMs were linear, PCA-based models of 3D shape variation, but several non-linear and deep learning-based extensions have been proposed during the last years [2, 6, 19, 64]. In the last decades, several 3D face models have been built from large datasets of 3D scans of human faces [7, 11, 15, 16, 31, 45, 50, 66, 69].

Monocular 3D Face Reconstruction: A common application of 3DMMs includes estimation of the model parameters that best fit to an RGB image. This can happen as a direct optimization procedure in an analysis-by-synthesis framework [5, 10, 12, 60, 62]. However this is a computationally expensive procedure to run on novel images every time. Due to this reason, various methods have emerged that formulate the problem as a regression from image data, leveraging the power of Deep Learning [22, 26, 30, 33, 39, 40, 51, 58, 63]. Combined with a reliable facial landmarker, this can lead to accurate results, even without the need for 3D supervision.

For example, RingNet [53], performs 3D reconstruction using the FLAME model [45], by enforcing a shape-consistency loss between images of the shape subject, to decouple identity and expression. This is improved by

DECA [25], which predicts the FLAME parameters jointly from a CNN, using multiple loss terms. EMOCA [21] focuses on the expressiveness of the reconstructed models, adding an emotion-related perceptual loss and training a CNN that predicts the expression parameters of the 3DMM on a large emotional dataset ExpNet [17] generates pseudo-3DMM parameters by solving the optimization problem given an accurate 3D reconstruction of an image with a SoTA method and then training a CNN to predict them, without the need for landmarks. In 3DDFA [34, 35, 71], face alignment and 3D reconstruction takes place concurrently, using Cascaded CNNs. MICA [72] focuses on accurate prediction of the identity parameters of a 3DMM, by employing a medium-scale 3D annotated dataset in conjunction with a large-scale 2D raw image dataset. DAD-3DHeads [48], provides one of the first large-scale 3D head datasets, that can be used for direct supervision of 3D reconstruction. Finally, most recently Wood et al. [68] used synthetic data for monocular 3D reconstruction which generalizes to real world footage. Some methods also try to deal with occlusions [23, 44, 54].

Even though the vast majority of methods reconstruct single face images or work on a frame-by-frame fashion on videos, there are a few methods that exploit the dynamic information of monocular face videos to constrain the subject's facial shape or impose temporal coherence on the face reconstruction [11, 14, 28, 38, 42].

A recent rising trend is exploiting deep features as metrics that correlate better with human perception compared to traditional metrics [70]. Our work is mostly similar to EMOCA [21], in the sense that both are concerned with perceptual reconstruction. In comparison, however, EMOCA focuses on retaining affective information from images while our work focuses on accurate reconstruction of mouth and lips formations that correspond to speech production. Furthermore, EMOCA failed to accurately predict the jaw pose (opening and rotation) of the mouth due to difficulties in convergence and kept the jaw pose fixed.

Mouth/Lip Reconstruction: Some of the earliest works focusing on the dynamics of mouth and lips for 3D reconstruction include the works of Basu et al. [8, 9] which used a combined-statistical model, Gregor et al. [41] who used markers to follow the lip motions, and Cheng et al. [18] who performed mouth tracking from 2D images using Adaboost and a Kalman filter. The most recent work concerned with lip tracking from video is the work of Garrido et al. [29], who achieved remarkable results of 3D reconstructed lips, using the ground truth shapes of a high quality 3D stereo database along with radial basis functions.

3. Proposed Method

Our work is based on the state-of-the-art DECA [25] framework for monocular 3D reconstruction from static RGB images. As such, we adopt the notation from the DECA paper. In the original DECA, given an input image I, a coarse encoder (a ResNet50 CNN) jointly predicts the identity parameters $\boldsymbol{\beta} \in \mathbb{R}^{100}$, neck pose and jaw $\boldsymbol{\theta} \in \mathbb{R}^{6}$, expression parameters $\boldsymbol{\psi} \in \mathbb{R}^{50}$, albedo $\boldsymbol{\alpha} \in \mathbb{R}^{50}$, lighting $\boldsymbol{I} \in \mathbb{R}^{27}$, and camera (scale and translation) $\boldsymbol{c} \in \mathbb{R}^{3}$. Note that these parameters are a subset of the parameters of the FLAME 3D face model. Afterwards, these parameters are used to render the predicted 3D face. DECA also included a detail encoder which predicted a latent vector associated with a UV-displacement map, that models high-frequency person-specific details such as wrinkles. More recently, EMOCA [21] further built upon DECA by adding an extra expression encoder (ResNet50) which was used in order to predict the expression vector $\boldsymbol{\psi}$, so that the perceived emotion of the reconstructed face is similar to that of the original image. We use these two works as starting points and design an architecture that improves the perceived expressions of the input video, concentrating on the mouth area, leading to realistic articulation movements.

3.1. Architecture

A high-level overview of the architecture is shown in Figure 2. Given a sequence of K RGB frames sampled from an input video V, our method reconstructs for each frame I the 3D mesh of the face in FLAME topology, such that the mouth movements and general facial expressions are perceptually preserved. Following the FLAME nomenclature, we separate the estimated parameters into two distinct sets:

Rigid & Identity parameters: We borrow the coarse encoder from DECA to predict independently for each image I in the input sequence the identity $\boldsymbol{\beta}$, neck pose $\boldsymbol{\theta_{neck}}$, albedo $\boldsymbol{\alpha} \in \mathbb{R}^{50}$, lighting $\boldsymbol{l} \in \mathbb{R}^{27}$, and camera $\boldsymbol{c}$. Like EMOCA [21], this network remains fixed through training.

Expression & Jaw parameters: The expression $\boldsymbol{\psi}$ and jaw pose $\boldsymbol{\theta}_{jaw}$ parameters that correspond to the input sequence is predicted by an additional "perceptual" CNN encoder, driven by deep perceptual losses that will be described shortly. These parameters explicitly control the mouth expressions and movements under the FLAME framework and therefore should be properly estimated by our approach. We employ a lightweight MobileNet v2 architecture, but also insert a temporal convolution kernel on its output, to model the temporal dynamics of mouth movements and facial expressions in the input. We selected the aforementioned lightweight option of MobileNet to reduce the computational overhead of our system - contrary to EMOCA- since the existing DECA backbone already uses a resource-demanding ResNet50 model.

In a nutshell, we assume an architecture akin to the one introduced in EMOCA [21], with two parallel paths of parameters as described above. Nevertheless, our focus is shifted to a very different problem and thus a set of ap-

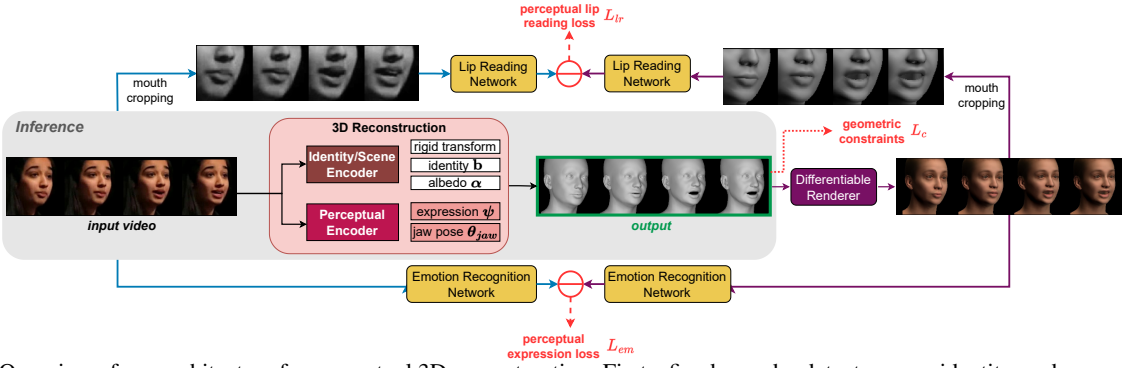

Figure 2. Overview of our architecture for perceptual 3D reconstruction. First a fixed encoder detects scene, identity, and a coarse estimate of jaw and expression. Then, a perceptual encoder refines the expression and jaw, and a differentiable renderer renders the predicted 3D shape (with texture). Finally, the input and rendered mouth are differentiably cropped and a pretrained lip reader is used to calculate the perceptual lip reading distance. Similarly, a perceptual expression loss is used on the full face based on an emotion recognition network. Inference requires only the 3D reconstruction component.

propriate "directions" and "constraints" should be learned through the use of the proposed set of losses, as described in the following section.

3.2. Training Losses

In order to train the *perceptual* encoder, we use two perceptual loss functions for guiding the reconstruction, along with geometric constraints. These losses are expected to correlate better with human perception (similarly to the loss functions based on deep features presented in [70]).

Perceptual Expression Loss: The output of the *perceptual* encoder is used along with the predictions of identity, albedo, camera, and lighting in order to differentiably render a sequence of textured 3D meshes, which correspond to the original input video. Then, the input video and the reconstructed 3D mesh are fed into an emotion recognition network (borrowed from EMOCA [21]) and two sequences of feature vectors are obtained. The perceptual expression loss L_{em} that we use corresponds to the distance between the two sequences of feature vectors. Interestingly, since the emotion recognition network is trained to predict emotions, it can faithfully retain a set of helpful facial characteristics. Therefore, such a loss is responsible for learning general facial expressions, also related to emotions that promote the realism of the derived reconstruction. Notably, this loss positively affects the eyes, leading to a more faithful estimation of eye closure, frowning actions, etc.

Perceptual Lip Movements Loss: The perceptual expression loss does not retain enough detailed information about the mouth, and as such, an additional mouth-related loss is needed. Instead of relying only on a geometric loss with weak supervision using 2D landmarks, we use an additional perceptual loss, that guides the output jaw and expression coefficients to capture the intricacies of mouth movements related to speech articulation. *The necessity of such a perceptual mouth-oriented loss is further highlighted by the inaccuracies detected in the extracted 2D landmarks* (please refer to the Suppl. Material for related examples).

For this purpose we use a network that has been trained on the LRS3 (Lip Reading in the Wild 3) dataset [46]. The lipreading network is the pretrained model provided by Ma et al. [46] that takes sequences of grayscale images cropped around the mouth as input and outputs the predicted character sequence. The network has been trained with a combination of Connectionist Temporal Classification (CTC) loss with attention. The model architecture consists of a 3D convolutional kernel, followed by a 2D ResNet-18, a 12-layer conformer, and finally a transformer decoder layer which outputs the predicted sequence (for more details, see [46]). Our goal here is to minimize the perceptual distance of speech-related movements between the original and the output image sequences. To that end, we take the differentiably rendered image sequences and subsequently crop them around the mouth area using the predicted landmarks. We calculate the corresponding feature vectors ϵ_I and ϵ_R, from the output of the 2D ResNet-18 of the lipreading network. We empirically found that these features better model the spatial structure of the mouth, while potentially alternative features based on the output of the conformer are largely influenced by the sequence context and do not preserve this much-needed spatial structure. Examples of this behavior can be found in the Suppl. Material. After calculating the feature vectors, we minimize the perceptual lip reading loss between the input image sequence and the output rendered sequence. The loss is defined as $L_{lr} = \frac{1}{K}\sum_{i=1}^{K} d(\epsilon_I^i, \epsilon_R^i)$, where d is the cosine distance and K the length of the input sequence. As a sidenote, initial experiments included an explicit lip reading loss based on the CTC loss over the predicted output of the existing lip reading network, given the original transcription of the sentence. However, this approach had major downsides apart from the need of the video transcription. First, it had a

significant computational overhead since whole sentences should be processed at once. In contrast, the proposed approach simply samples a subset of consecutive frames and tries to minimize the extracted mouth-related features. Moreover, it was proven ineffective in practice, suffering from the same behavior as with the conformer's features.

Geometric Constraints: Due to the domain mismatch between the rendered and the original images, although the perceptual losses help retain the high level information on perception, they also tend to create artifacts in some cases. This is to be expected; the perceptual losses rely on pretrained task-specific CNNs that do not guarantee in any way that the input manifold corresponds to realistic images. For example, as we report in Suppl. Material, we can create unrealistic images of distorted facial reconstruction that produce good lip reading results - a typical problem in the adversarial examples topic [32]. Thus, we guide the training process by enforcing the following geometric constraints: We regularize the expression and jaw parameters by penalizing their L_2 norm: $||\boldsymbol{\psi}||^2$ and $||\boldsymbol{\theta}_{jaw}||^2$. We also apply an L_1 loss (average per-landmark distance) between the landmarks of the nose, face outline and eyes of the 3D model and the predicted landmarks of a face alignment method [13]. For the mouth area we employ a more relaxed L_2 relative loss between the intra-distances of mouth landmarks. The aforementioned landmark losses comprise an alternative to explicitly imposing a geometric loss based on distance between the predicted 2D landmarks of the reconstructed face and the 2D landmarks of the original image. Such a straightforward loss can lead to erroneous reconstruction, as the ablation study in Suppl. Material highlights, since perceptual losses and the 2D landmark loss were often contradicting. Using the proposed version of relative landmark losses achieves retaining the much needed geometric structure of the face without an overly strict constraint that limits the perceptual losses.

Finally, the total loss used for training is then: $L = \lambda_{lr} L_{lr} + \lambda_{em} L_{em} + Lc$, where L_c includes the previously stated geometric constraints.

4. Experiments

We evaluate our method both qualitatively and quantitatively, following a similar evaluation procedure to [21]. The considered datasets are the following:

• **LRS3** [3]: We use Lip Reading Sentences 3 (LRS3) dataset [3], the largest publicly available dataset for lip reading in the wild, for training and testing our system. The official *trainval* set (31,982 utterances) is used for training and validating our model, while evaluation is performed on the test set of LRS3 (1,321 utterances).

• **MEAD** [65]: This is a recent dataset containing 48 actors (28M, 20F) from multiple races uttering sentences from TIMIT [27] in 7 emotions and 3 different levels of intensity. The whole dataset includes 31,059 sentences. We randomly sampled 2,000 in order to create a test set, stratifying for subject, emotion, and intensity level.

• **TCD-TIMIT** [36]: This corpus includes 62 English actors reading 6913 sentences from the TIMIT [27] corpus. We use the official test split for evaluation.

• **VOCASET** [20]: VOCASET includes 12 subjects speaking 40 utterances each. It is the only dataset which includes ground truth registered vertices in the FLAME mesh topology, enabling evaluation with geometric-based metrics. We use the official test split for evaluation.

Training Details: We follow a two-stage training scheme using Adam optimizer with batch size 1 and sequence length $K = 20$. Source code and more details on the training procedure are provided in the Suppl. Material.

Comparisons: We compare our method to the following recent state-of-the-art methods on 3D facial reconstruction: **DECA** [25], **EMOCA** [21], **3DDFAv2** [34], and **DAD-3DHeads**, which uses direct 3D supervision from the large-scale annotated DAD-3DHeads [48] dataset. Note that these methods, as almost all recent methods for visual reconstruction of the 3D face geometry, are using a single RGB image as input. Therefore, in order to reconstruct the entire input video, we apply them in every frame of the video. Especially for 3DDFAv2, we apply temporal smoothing as provided by the official implementation. For all methods we use the official implementation. In Fig. 3 multiple visual comparisons with the other methods can be seen. **Additional results are provided in the Suppl. Material.**

4.1. Quantitative Evaluation

In this section, we seek to quantify speech-related perceptual cues. A straightforward way is to evaluate the compared methods objectively in terms of lip reading metrics by applying a pretrained lipreading network on the output rendered images. To remove bias, we use a *different architecture and pretrained lipreading model* for evaluation than the one used for the lipreading loss, which is based on the Hubert transformer architecture, called AV-HuBERT [55, 56]. The following lipreading metrics are considered: Character Error Rate (CER), and Viseme Error Rate (VER) (obtained by converting the predicted and ground truth transcriptions to visemes using the Amazon Polly phoneme-to-viseme mapping [1]) when using greedy decoding, as well as the per-frame accuracy of AV-HuBERT, which predicts for each frame one from 1000 subword classes.

We first present results on the VOCASET dataset, which contains ground truth 3D reconstruction, in Table 1. Apart from the CER and VER, we also report the $L2$ per-vertex error in mm, for the mouth, non-mouth and full face regions. Note how lipreading results are not correlated with the latter set of geometric errors/scores. Specifically, our approach leads to significantly improved lipreading metrics

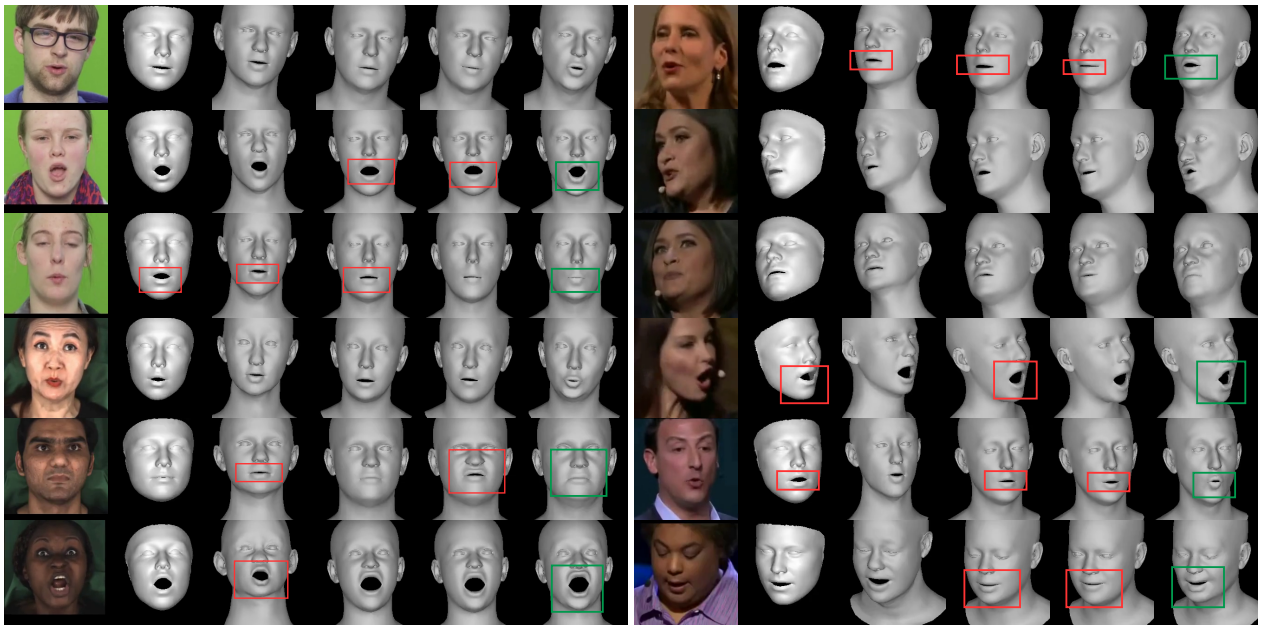

Figure 3. Visual comparison with other methods on the MEAD, TCDTIMIT, and LRS3 datasets. Note that our method is only trained on the LRS3 train test. From left to right: original footage, 3DDFAv2 [34], DAD [48], DECA [25], EMOCA [21], ours. We also highlight with red boxes some erroneous results, and with green boxes some examples of retaining the original mouth formation.

	CER ↓	VER ↓	mouth L2 ↓	non-mouth L2 ↓	full L2 ↓
Orig.	42.6	32.6	-	-	-
DECA	100.9	89.9	5.92	5.14	5.25
EMOCA	100.9	90.7	6.64	5.19	5.39
DAD	92.3	86.6	**4.07**	**4.89**	**4.78**
SPECTRE	**87.6**	**77.0**	5.39	5.55	5.56

Table 1. Lipreading (CER, VER) and geometric metrics on vertices (mouth, nonmouth, and full face $L2$) are reported on the VOCASET test set. While SPECTRE achieves significantly better lipreading metrics, this result is not reflected on traditional geometric errors ($L2$ scaled by $\times 10^3$).

compared to the other reconstruction methods, as expected, while DAD achieves the best $L2$ scores, powered by its detailed 3D supervision approach. Compared to DECA and EMOCA, our method achieves a better mouth $L2$ value despite being trained on the proposed lipreading loss and thus expecting loose geometric correspondence. For further validation, we show two example snapshots from VOCASET in Fig. 4 where it can be seen that the per-vertex geometric error does not represent well the mouth formation and is not representative of the perceived 3D reconstruction quality. This result has been highlighted by various previous works, which have pointed out that geometric errors of facial/mouth expressions do not correlate well with human perception [4, 21, 29, 49, 59, 67].

For the rest of datasets we do not have ground truth landmarks, and predicted ones tend to not capture mouth formations well (see Suppl. Material for examples). As a result, we evaluate only the aforementioned lipreading metrics.

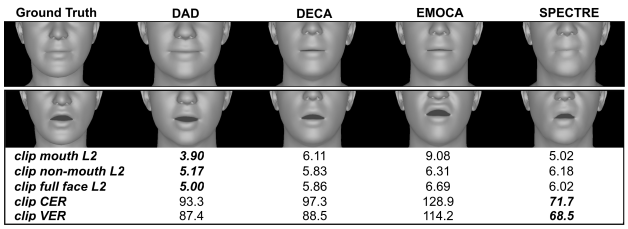

Figure 4. Comparison of mouth area 3D reconstructions from a VOCASET clip with reported L2, CER, and VER errors (best results in bold). L2 errors are scaled by $\times 10^3$. Notice the discrepancy in the ranking of the different methods between $L2$ metrics and CER/VER metrics. We observe that the perceived quality of mouth reconstruction seems to have a much better correlation with CER and VER metrics, rather than $L2$. For better inspection of the results, please zoom in and refer to the Suppl. video.

First, we show results for CER and VER in Table 2. Our method achieves considerably lower CER and VER scores compared to the other methods, both in the LRS3 test set, as well as in the cross-dataset evaluations of TCDTIMIT and MEAD. In the same Table we also include results on the original video footage, which showcase the domain gap "problem" (more information about this in Discussion section) of the used lip reading systems: the pre-trained models have been trained to the initial images without the possible visual degradation introduced by the rendering procedure.

Fig. 5 also shows a detailed lipread analysis on all datasets using top-k accuracy curve for varying k (across 1000 AV-HuBERT classes). SPECTRE's curve is con-

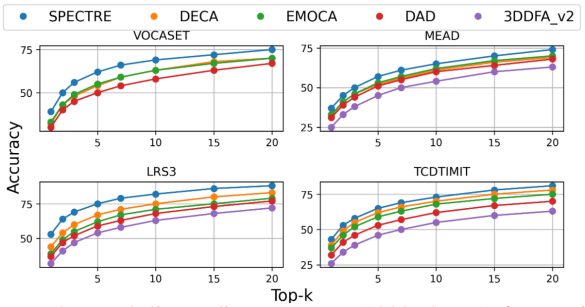

Figure 5. Top-k lip reading accuracy (1000 classes) for varying k in VOCASET, MEAD, LRS3 and TCDTIMIT datasets. SPECTRE's curve is consistently above other methods.

sistently higher, confirming its lip modeling performance. This boost is achieved despite missing features like tongue and teeth, by effectively encoding speech-related features.

	LRS3		TCD-TIMIT		MEAD	
	CER ↓	VER ↓	CER ↓	VER ↓	CER ↓	VER ↓
Original vid	24.9	22.0	35.7	29.6	49.7	42.8
DECA	77.5	70.8	84.2	75.8	84.8	77.8
EMOCA	83.3	76.3	86.4	79.2	85.1	77.9
3DDFAv2	97.5	95.3	101.8	98	94.5	90.2
DAD	84.1	78.2	87.3	81	86.0	79.9
SPECTRE	**67.5**	**60.9**	**78.1**	**69.6**	**78.5**	**71.1**

Table 2. Lipreading results on the LRS3-test, TCD-TIMIT and MEAD datasets (network trained on LRS3-train set). Our method significantly outperforms all other 3D reconstruction methods.

4.2. User Studies

The quantitative evaluation highlighted the difficulty to pin down well-received perceptual cues into a concrete geometric error. In fact, introducing a realistic, non-excessive over-articulation should be favorable with respect to human perception despite the expected deviation from geometric errors, as pointed out in [4]. Arguably, the ultimate goal of a talking head is for humans to perceive it as natural and as realistic as possible. To assess the realism and perception of the 3D reconstructed faces by humans, we have designed and conducted two web user studies [43]. In order to mitigate any intra-dataset bias that might arise from training on the LRS3 trainset, for these studies, we used only videos from the MEAD and TCD-TIMIT datasets.

First Study: Realism of Articulation. For this study, we selected a preference test design, by showing users pairs of videos with 3D face reconstruction results, alongside the original footage, and asking them to select the most realistic one in terms of mouth movements and articulation. We created a question bank consisting of 30 videos from the MEAD dataset (21 emotional videos for each level of intensity and emotion and 9 neutral), and 10 videos from the TCD-TIMIT dataset and performed 3D reconstruction using the previously stated 5 methods (DAD, DECA, EMOCA, 3DDFAv2 and ours). Then, users were presented with two randomly ordered reconstructed faces, alongside the original footage, and were asked to choose the most realistic one in terms of mouth movements and articulation. Each user answered 28 randomly sampled questions from the bank (7 questions for each pair - ours vs the others), and a total of 34 users completed this study.

The results of this study can be seen in Table 3. We can see that our method is significantly preferred to all other methods ($p < 0.01$ with binomial test, adjusting for multiple comparisons using the Bonferroni method). 3DDFAv2 [34] was the least preferred method, with DECA and EMOCA following. The results clearly highlight the importance of the proposed method from the speech-aware perspective and how humans favorably perceive the reconstructed mouth movements as more realistic in SPECTRE, compared to the other methods.

	DECA	EMOCA	3DDFAv2	DAD
SPECTRE	**201**/37	**185**/53	**218**/20	**150**/88

Table 3. First subjective study preference results: "a/b" indicates SPECTRE (left) was preferred a times, while the competing method was chosen b times (238 pairs assessed). SPECTRE is significantly more realistic in mouth movements and articulation.

Second Study: Lip Reading. In the second study, a different set of users were presented with a muted video of a person uttering a single word in the form of a 3D talking head reconstructed from one of the compared methods and then were asked to select the correct word among 4 different alternatives (multiple choice). For this, we cropped 40 single words from the MEAD and TCD-TIMIT datasets, covering different visemes, and presented each user with a random subset of 30 words (6 words for each method in each questionnaire). A total of 31 users completed this study. Classification results are shown in Table 4. In a similar fashion with the objective results, SPECTRE outperforms other methods in terms of word classification. An interesting result is the fact that EMOCA achieves a relatively high result compared to objective results. This could be due to the fact that in some cases, e.g., unrealistically exaggerated expressions as seen in EMOCA, can be sufficient for distinguish specific words. A per word analysis with visual examples is also provided in the Suppl. Material.

SPECTRE	DECA	EMOCA	3DDFAv2	DAD
47.56%	39.83%	45.12%	23.17%	45.12%

Table 4. Word classification accuracy in the second user study.

4.3. Visual Comparisons and Ablation

We conduct an ablation study on the effect of the temporal convolution and the lip reading loss. First in Tab. 5 we show an $L2$-based evaluation of the 3D rec. quality on VOCASET. Without the lipread loss, mouth $L2$ is much higher,

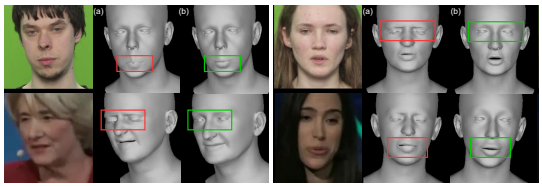

Figure 6. Training without (a) and with (b) geometric constraints. Lack of constraints causes artifacts in eyes/nose, while excluding mouth landmarks may lead to failures in mouth area.

since the geometric constraints include only the relative positions of mouth landmarks. Omitting the temporal convolution degrades results too, since the temporal dynamics of the lipread network (which also uses temporal convolutions) are not properly accounted for. Second, in Tab. 6 we also show the results of a user study (with 22 users) on the perceived realism when omitting these components. We see that users significantly preferred the full model, compared to the models that did not employ temporal convolution or the lipread loss. Finally, in Fig. 6 we also show results of training the network with and without the geometric constraints from landmarks. In some cases, removing geometric constraints and training only with perceptual losses leads to artifacts around the eyes, nose and mouth shape. For more ablation studies, see the Suppl. Material.

temporal conv.	lipread loss	mouth $L2$	non-mouth $L2$	full $L2$
✓	✗	7.49	**5.33**	5.63
✗	✓	8.69	5.56	6.01
✓	✓	**5.39**	5.55	**5.56**

Table 5. Ablation study on the effect of removing the temporal convolution or the lip reading loss from SPECTRE.

5. Discussion

We performed an important step towards perceptually realistic 3D talking heads, as shown by our extensive evaluations against other SoTA methods. Notably, our method even outperforms DAD in terms of realism, which was trained with 3D annotated data on a large-scale dataset. Even though DAD was shown to achieve a geometrically accurate 3D shape, the lack of perceptual losses rendered the result less realistic, compared to SPECTRE. It should also be pointed out (also seen in Figure 3), that the lipread loss, not only retains the motions and shape of the mouth, but it also makes it more distinct in the rendered mesh. It becomes apparent that in order to achieve realism in terms of speech, we need to opt for more perceptual losses. This has also been done in previous methods regarding emotional expression [21] as well as 3D shape [25, 72]. SPECTRE also has various applications, including audio/text-driven talking heads. It enables acquiring high-quality 3D data from in-the-wild videos, bypassing time-consuming 3D data collection for audio/text-driven training. Alternatively, the lipread loss can be directly used for training other models.

	ours w/o temporal conv	ours w/o lipread
Ours	**97**/35	**98**/34

Table 6. Ablation user study on the effect of removing the temporal convolution or the lip reading loss from SPECTRE.

Limitations We point out that the objective evaluation results on CER and WER, remain much higher compared to the original footage. This can be attributed to the different domains of the rendered images compared to the ground truth, as well as the absence of teeth and tongue, which are important for detecting specific types of phonemes/visemes. This domain adaptation problem is not fully addressed in this work, since our approach is effective in practice, but it remains a hindrance to unleashing the full potential of the described perceptual losses, which are also affected. These losses they assume that the original images and the rendered ones belong to the same visual "domain". Nonetheless, this domain gap between these two feature spaces may lead to inconsistencies; this is why we needed to have relative landmarks. As a result, the geometric loss and the lipreading loss sometimes compete against each other: on one hand, lip reading tries to improve the perception of the talking head while landmarks, if not detected accurately, tend to reduce the realism. On the other hand, we observed that below a certain threshold, reduction of lip reading loss tends to create artifacts; which is why we needed the constrains from landmarks to retain the realism of the facial shape. Also, while our method includes an emotion recognition loss ([21]), since it was trained only on the LRS3 dataset (which does not include emotional samples) results tend to not achieve the emotional intensity in EMOCA. Finally, while SPECTRE does not use text or audio, these modalities could be leveraged in order to improve the total perception, or, bypass problems such as visual occlusions.

6. Conclusion

We presented the first method for visual speech-informed perceptual 3D reconstruction of talking heads. Our method does not use text transcriptions or audio but employs a "lipreading" loss, to increase mouth perception. Our extensive evaluations verified that our method is significantly preferred to counterparts which rely only on geometric losses for mouth movements, as well as to ones that use direct 3D supervision. This is an important step towards reconstructing perceptually realistic talking heads, by focusing not only on the purely geometric-based aspects, but also on human perception of speech articulation.

Acknowledgments. This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101070381 (project: PILLAR-Robots).

References

[1] Amazon Polly. Developer Guide., 2015. 5

[2] Victoria Fernández Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, and Edmond Boyer. A generative 3d facial model by adversarial training. 2019. 2

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 5

[4] Zakaria Aldeneh, Masha Fedzechkina, Skyler Seto, Katherine Metcalf, Miguel Sarabia, Nicholas Apostoloff, and Barry-John Theobald. Towards a Perceptual Model for Estimating the Quality of Visual Speech, Mar. 2022. arXiv:2203.10117 [cs, eess]. 6, 7

[5] Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3d morphable model. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1080–1093, 2012. 2

[6] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 2

[7] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, Dong Yu, and Zhengyou Zhang. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics*, 2021. 2

[8] Sumit Basu, Nuria Oliver, and Alex Pentland. 3d lip shapes from video: A combined physical–statistical model. *Speech Communication*, 26(1-2):131–148, 1998. 3

[9] Sumit Basu, Nuria Oliver, and Alex Pentland. 3d modeling and tracking of human lip motions. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 337–343. IEEE, 1998. 3

[10] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 1999. 2

[11] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018. 2, 3

[12] James Booth, Anastasios Roussos, Evangelos Ververas, Epameinondas Antonakos, Stylianos Ploumpis, Yannis Panagakis, and Stefanos Zafeiriou. 3d reconstruction of "in-the-wild" faces in images and videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2638–2652, 2018. 2

[13] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 5

[14] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)*, 34(4):1–9, 2015. 3

[15] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 2

[16] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[17] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 122–129. IEEE, 2018. 3

[18] Jie Cheng and Peisen Huang. Real-time mouth tracking and 3d reconstruction. In *2010 3rd International Congress on Image and Signal Processing*, volume 4, pages 1524–1528. IEEE, 2010. 3

[19] Shiyang Cheng, Michael Bronstein, Yuxiang Zhou, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. Meshgan: Non-linear 3d morphable models of faces. *arXiv preprint arXiv:1903.10384*, 2019. 2

[20] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 5

[21] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2, 3, 4, 5, 6, 8

[22] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 285–295, Long Beach, CA, USA, June 2019. IEEE. 2

[23] Rahul Dey and Vishnu Naresh Boddeti. Generating Diverse 3D Reconstructions from a Single Occluded Face Image. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1537–1547, New Orleans, LA, USA, June 2022. IEEE. 3

[24] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models—past, present, and future. *ACM Trans. Graph.*, 39(5), 2020. 2

[25] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 2, 3, 5, 6, 8

[26] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 2

[27] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-

phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403, 1993. 5

[28] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):1–15, 2016. 3

[29] Pablo Garrido, Michael Zollhöfer, Chenglei Wu, Derek Bradley, Patrick Pérez, Thabo Beeler, and Christian Theobalt. Corrective 3d reconstruction of lips from monocular video. *ACM Trans. Graph.*, 35(6):219–1, 2016. 3, 6

[30] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1155–1164, 2019. 2

[31] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schoenborn, and Thomas Vetter. Morphable face models - an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82, 2018. 2

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 5

[33] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. *arXiv preprint arXiv:2112.01554*, 2021. 2

[34] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3ddfa. https://github.com/cleardusk/3DDFA, 2018. 3, 5, 6, 7

[35] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[36] Naomi Harte and Eoin Gillen. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015. 5

[37] Matthias Hofer, Tilo Hartmann, Allison Eden, Rabindra Ratan, and Lindsay Hahn. The role of plausibility in the experience of spatial presence in virtual environments. *Frontiers in Virtual Reality*, page 2, 2020. 2

[38] Patrik Huber, Philipp Kopp, William Christmas, Matthias Rätsch, and Josef Kittler. Real-time 3d face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*, 24(4):437–441, 2016. 3

[39] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pages 1031–1039, 2017. 2

[40] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016. 2

[41] Gregor A. Kalberer and Luc J. Van Gool. Lip animation based on observed 3D speech dynamics. In Sabry F. El-Hakim and Armin Gruen, editors, *Videometrics and Optical Methods for 3D Shape Measurement*, volume 4309, pages 16 – 25. International Society for Optics and Photonics, SPIE, 2000. 3

[42] Mohammad Rami Koujan and Anastasios Roussos. Combining dense nonrigid structure from motion and 3d morphable models for monocular 4d face reconstruction. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–9, 2018. 3

[43] Kosmas Kritsis, Aggelos Gkiokas, Aggelos Pikrakis, and Vassilis Katsouros. Danceconv: Dance motion generation with convolutional networks. *IEEE Access*, 10:44982–45000, 2022. 7

[44] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. *arXiv preprint arXiv:2106.09614*, 2021. 3

[45] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2

[46] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual Speech Recognition for Multiple Languages in the Wild. *arXiv Preprint: 2202.13084*, 2022. 4

[47] Javier Marín-Morales, Carmen Llinares, Jaime Guixeres, and Mariano Alcañiz. Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors*, 20(18):5163, 2020. 2

[48] Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiři Matas, and Viktoriia Sharmanska. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5, 6

[49] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012. 6

[50] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2

[51] Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE Transactions on Image Processing*, 30:5793–5806, 2021. 2

[52] Shunsuke Saito, Tianye Li, and Hao Li. Real-time facial segmentation and performance capture from rgb input. In *European conference on computer vision*, pages 244–261. Springer, 2016. 2

[53] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings IEEE Conf.*

on Computer Vision and Pattern Recognition (CVPR), June 2019. 2

[54] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*, 2020. 3

[55] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 5

[56] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022. 5

[57] Jacob Stuart, Karen Aul, Anita Stephen, Michael D Bumbach, and Benjamin Lok. The effect of virtual human rendering style on user perceptions of visual cues. *Frontiers in Virtual Reality*, page 58, 2022. 2

[58] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[59] Barry-John Theobald and Iain Matthews. Relating Objective and Subjective Performance Measures for AAM-Based Visual Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2378–2387, Oct. 2012. 6

[60] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6), 2015. 2

[61] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *arXiv preprint arXiv:1610.03151*, 2016. 2

[62] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *Commun. ACM*, 62(1):96–104, 2018. 2

[63] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018. 2

[64] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2

[65] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 5

[66] Lizhen Wang, Zhiyua Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2022)*, June 2022. 2

[67] Danny Websdale, Sarah Taylor, and Ben Milner. Speaker-Independent Speech Animation Using Perceptual Loss Functions and Synthetic Data. *IEEE Transactions on Multimedia*, 24:2539–2552, 2022. 6

[68] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3d face reconstruction with dense landmarks. In *Proc. ECCV*, 2022. 3

[69] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3, 4

[71] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 3

[72] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. 2022. 3, 8

[73] Michael Zollhöfer, Justus Thies, Darek Bradley, Pablo Garrido, Thabo Beeler, Patrick Péerez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. 2018. 2