Self-Supervised 3D Face Reconstruction via Conditional Estimation

Yandong Wen[1] Weiyang Liu[2,3] Bhiksha Raj[1] Rita Singh[1]
[1]Carnegie Mellon University [2]University of Cambridge [3]MPI for Intelligent Systems, Tübingen

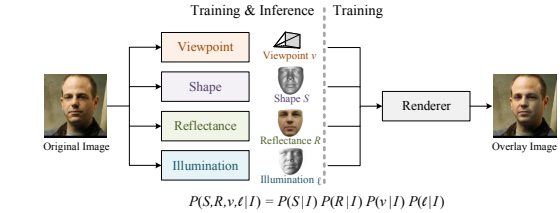
Abstract

We present a conditional estimation (CEST) framework to learn 3D facial parameters from 2D single-view images by self-supervised training from videos. CEST is based on the process of analysis by synthesis, where the 3D facial parameters (shape, reflectance, viewpoint, and illumination) are estimated from the face image, and then recombined to reconstruct the 2D face image. In order to learn semantically meaningful 3D facial parameters without explicit access to their labels, CEST couples the estimation of different 3D facial parameters by taking their statistical dependency into account. Specifically, the estimation of any 3D facial parameter is not only conditioned on the given image, but also on the facial parameters that have already been derived. Moreover, the reflectance symmetry and consistency among the video frames are adopted to improve the disentanglement of facial parameters. Together with a novel strategy for incorporating the reflectance symmetry and consistency, CEST can be efficiently trained with in-the-wild video clips. Both qualitative and quantitative experiments demonstrate the effectiveness of CEST.

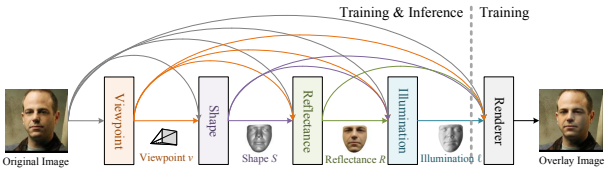
1. Introduction

Reconstructing 3D faces from single-view 2D images has been a longstanding problem in computer vision. The common approach represents the 3D face as a combination of its *shape*, as represented by the 3D coordinates of a number of points on its surface called vertices, and its *texture*, as represented by the reflectances of red, green and blue at these vertices [4]. The problem then becomes learning a regression model between the 2D images, and vertices and their reflectances.

The regression itself may be learned using training data where both, the 2D images and the corresponding 3D parameters are available. However, these data are scarce, and even the ones that are available generally only have shape information [8, 47, 46]; the ones that do have other parameters are usually captured in a controlled environment [22] or are synthetic [33], which is not representative of real-world images. Consequently, there is great interest in self-supervised learning methods, which learn the regression model from natural in-the-wild 2D images or videos, without explicit access to 3D training data [39, 41].



(a) Most existing methods assume conditional independence.



(b) The proposed CEST makes no assumption on dependency.

Figure 1: Conventional 3D face reconstruction and our CEST framework. The dotted lines separate the modules used for inference of the 3D parameters from those used for training with self-supervision.

The problem is complicated by the fact that the actual image formation depends not only on the shape and texture of the face, but also the illumination (the intensity and direction of the incident light), and other factors such as the viewpoint (incorporating the orientation of the face and the position of the camera), etc. Thus, the learned regression model must also account for these factors. To this end, the general approach is one where shape, reflectance, illumination and viewpoint parameters are all extracted from the 2D image. The regression model that extracts these facial parameters are learned through *self-supervision*: the extracted facial parameters are recombined to render the original 2D image, and the model parameters are learned to minimize the reconstruction error.

The solution, however, remains ambiguous because a 2D image may be obtained from different combinations of shape, texture, illumination and viewpoint. To ensure that the self-supervision provides meaningful disentanglement, the manner in which the facial parameters are recombined to reconstruct the 2D image are based on the actual physics of image formation [39, 41, 33]. To further reduce potential ambiguities, regularizations are necessary. Reflectance *symmetry* has been proposed as a regularizer [42, 38, 45], wherein the reflectance of a face image and its mirror reflec-

tion are assumed to be identical. Smoothness has also been employed to regularize the shape and reflectance [41, 38]. Additional regularization may be obtained by considering correspondences between multiple images of the same face [18, 37], particularly when they are obtained under near identical conditions such as the sequence of images from a video. The approach in [37] has considered reflectance *consistency*, where reflectances of all image frames in a video clip are assumed to be similar.

In all of these prior works, the target parameters, namely the shape, reflectance, illumination and viewpoint parameters are all *individually* estimated, without considering their direct influences on one another, although they are jointly optimized. In effect, at inference time they assume that the estimate of, *e.g.* the reflectance, is conditionally independent of the estimated shape or viewpoint, given the original 2D image. The coupling among the four is only considered during (self-supervised) training, where they must all combine to faithfully recreate the input 2D image [11, 14, 29, 42, 37]. This is illustrated in Fig. 1(a).

In reality, 2D images are reduced-dimensional projections, and thus imperfect representations of the full three-dimensional structure of the face, and the aspects of reflectance and illumination imprinted in them are not independent of the underlying shape of the object or the viewpoint they were captured from. Therefore, the captured 2D image represents a joint interaction among viewpoint, shape, reflectance and illumination. Consequently, the statistical estimates of any of these four factors may not, in fact, be truly conditionally independent of one another given only the 2D image (although, given the entire 3D model they might have been). Thus, modelling all of these variables as being conditionally independent effectively represents a lost opportunity since, by predicting them individually, the constraints they impose on one another are ignored. Optimization-based approaches [17, 18, 35] do attempt to capture the dependence by iteratively estimating shape and reflectance from one another. However, these methods require correspondence information of the image sequence in a video and suffer from costly inference.

In this paper, we propose a novel learning-based framework based on conditional estimation (CEST). CEST explicitly considers the statistical dependency of the various 3D facial parameters (shape, viewpoint, reflectance and illumination) upon one another, when derived from single 2D image. The specific form of the dependencies adopted in this paper is shown in Fig. 1(b). We note that the CEST framework is very general and allows us to consider any other dependency structures. Our paper serves as one of the many potential choices that work well in practice. To this end, we present a specific, and intuitive, solution in CEST, where the viewpoint, facial shape, facial reflectance, and illumination are predicted *sequentially* and *conditionally*. In this context, the prediction of facial shape is conditioned on the input image and the derived viewpoint; the prediction of facial reflectance is conditioned on the input image, derived viewpoint and facial shape; and so forth.

As before, learning remains self-supervised, through comparison of re-rendered 2D images obtained with the estimated 3D face parameters to the original images. As additional regularizers, we also employ reflectance symmetry constraints [42, 38, 45], and reflectance consistency constraints (across frames in a short video clip) [37]. These are included in the form of cross-frame reconstruction error terms, the number of which increases quadratically with the number of video frames considered together for self-supervision. To address the dramatically increased number of reconstruction terms, we propose a stochastic optimization strategy to improve training efficiency.

We present ablation studies and comparisons to state-of-the-art methods [39, 42, 37] to evaluate CEST. We show that CEST produces better reflectance and structured illumination, leading to more realistic rendered faces with fine facial details, compared to all other tested methods. It also achieves better shape estimation accuracy on AFLW2000-3D [49] and MICC [1] datasets than current state-of-the-art self-supervised and *fully supervised* approaches. Overall, our contributions can be summarized as follows:

- We propose CEST, a conditional estimation framework for 3D face reconstruction that explicitly considers the statistical dependencies among 3D face parameters.
- We propose a specific design for the decomposition of conditional estimation, where the viewpoint, shape, reflectance, and the illumination are derived sequentially.
- We propose a stochastic optimization strategy to efficiently incorporate reflectance symmetry and consistency constraints into CEST. As the number of video frames increase, the computational complexity of CEST is increased linearly, rather than quadratically.

2. Related Work

Monocular 3D face reconstruction by self-supervised learning. Many research studies published recently aim to learn 3D facial parameters from a single image in a self-supervised manner. In [29], the authors propose a coarse-to-fine framework to improve the details in reconstructed 3D faces. Ayush *et al.* [39] present a model-based deep convolutional face autoencoder (MoFA) to fit a 3DMM to shape, reflectance, and illuminance. InverseFaceNet [20] trains a direct regression model on a synthetic training corpus that is generated by self-supervised bootstrapping. SfSNet [33] combines labeled synthetic and unlabeled real-world images in learning, and produces accurate depth map, and reflectance and shade disentanglement. To better characterize facial details, 3DMM is generalized to a nonlinear model in [41, 42]. [48] uses mesh convolutions for 3D faces, lead-

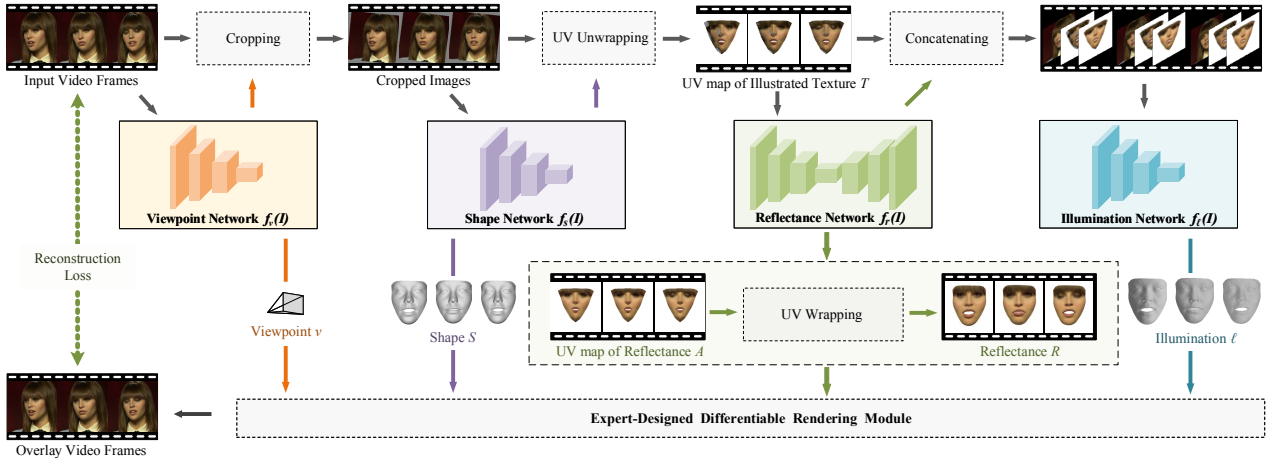


Figure 2: The overall training pipeline of the proposed CEST framework.

ing to a light-weight model with competitive performance. [34] incorporates the multi-view consistency from geometry, pixel, and depth as constraints.

However, these approaches generally do not consider correspondences across frames in a video. FML [37] is the first self-supervised framework that incorporates video clues in training. The shape and reflectance for each video frame are approximated by averaging the shapes and reflectances in a video clip. However, models trained on the averaged representations may not work well for a single image if the number of multi-frame images is large, due to the large gap between averaged and isolated images. On the contrary, CEST uses representations from single images. More importantly, it uses conditional estimation for predicting the facial parameters, and does not assume conditional independence between them, an often unrealistic assumption employed in the previously mentioned approaches.

Optimization-based 3D face reconstruction. [18] proposes to fit a template model to photo-collections by updating the viewpoint, geometry, lighting, and texture iteratively. [35] fits a face model to detected 3D landmarks, and refines the texture and geometry details. [11] learns facial subspaces for identity and expression variations with a parametric shape prior. [10] considers 3D face reconstruction as a global variational energy minimization problem, and estimates dense low-rank 3D shapes for video frames.

While these approaches can be considered conditional estimation, they focus on deriving 3D facial parameters from video, and are not relevant to the problem of deriving them from single-frame images, the problem addressed in our work. For CEST, video clips are viewed as consistent collections of images used to better learn the model.

3. The CEST Framework

In this work, we adopt a common practice from 3D Morphable Model (3DMM) [4], which represents a 3D face as a combination of shape and reflectance. The shape comprises a collection of vertices $\boldsymbol{S} = [\boldsymbol{S}(1); \boldsymbol{S}(2); ...; \boldsymbol{S}(K)] \in \mathbb{R}^{K\times 3}$, where K is the number of vertices and $\boldsymbol{S}(i) = [\boldsymbol{S}(i,1), \boldsymbol{S}(i,2), \boldsymbol{S}(i,3)]$ denotes the xyz coordinates in the Cartesian coordinate system. The typology for $\boldsymbol{S}$ is consistent for different faces. The reflectance comprises a collection of pixel values $\boldsymbol{R} = [\boldsymbol{R}(1); \boldsymbol{R}(2); ...; \boldsymbol{R}(K)] \in \mathbb{R}^{K\times 3}$. Each row $\boldsymbol{R}(i) = [\boldsymbol{R}(i,1), \boldsymbol{R}(i,2), \boldsymbol{R}(i,3)]$ comprises the pixel values (*i.e.*, RGB) at position $\boldsymbol{S}(i)$.

3.1. Framework Overview

The problem of 3D face reconstruction from a 2D image is that of obtaining estimates of the shape $\boldsymbol{S}$, reflectance $\boldsymbol{R}$, viewpoint $\boldsymbol{v}$ and illumination $\boldsymbol{\ell}$, given an input image $\boldsymbol{I}$. Statistically, we aim to estimate the most likely values for these variables, given the input image:

$$\hat{\boldsymbol{S}}, \hat{\boldsymbol{R}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{\ell}} = \arg\max_{\boldsymbol{S},\boldsymbol{R},\boldsymbol{v},\boldsymbol{\ell}} P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I}) \tag{1}$$

The challenges of this estimation are twofold: first $P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I})$ must be modelled, and second, $\arg\max_{\boldsymbol{S},\boldsymbol{R},\boldsymbol{v},\boldsymbol{\ell}} P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I})$ must be computed.

Modelling $P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I})$ directly is a challenging problem, and the problem must be factored down. Prior approaches [41, 39, 48] have decomposed this problem by assuming that shape, reflectance, viewpoint and illumination are all conditionally independent, given the image, i.e. $P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I}) = P(\boldsymbol{S}|\boldsymbol{I})P(\boldsymbol{R}|\boldsymbol{I})P(\boldsymbol{v}|\boldsymbol{I})P(\boldsymbol{\ell}|\boldsymbol{I})$. This leads to simplified estimates where each of the variables can be independently estimated, i.e. $\hat{\boldsymbol{S}} = \arg\max_{\boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{I})$, $\hat{\boldsymbol{R}} = \arg\max_{\boldsymbol{R}} P(\boldsymbol{R}|\boldsymbol{I})$, etc. As we have discussed earlier, the conditional independence assumption is questionable, since the conditioning variable, $\boldsymbol{I}$, is a lower-dimensional projection of the 3D face that entangles the four variables.

In CEST we explicitly model the conditional dependence, as shown in Fig. 1(b). Specifically we decompose the joint probability as

$$\begin{aligned} &P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|I) \\ &= P(\boldsymbol{v}|\boldsymbol{I})P(\boldsymbol{S}|\boldsymbol{I}, \boldsymbol{v})P(\boldsymbol{R}|\boldsymbol{I}, \boldsymbol{v}, \boldsymbol{S})P(\boldsymbol{\ell}|\boldsymbol{I}, \boldsymbol{v}, \boldsymbol{S}, \boldsymbol{R}) \end{aligned} \tag{2}$$

Coupling the variables in this manner results in a complication: even factored as above, maximizing the joint probability with respect to $\boldsymbol{S}$, $\boldsymbol{R}$, $\boldsymbol{v}$, and $\boldsymbol{\ell}$ must be jointly performed, since the variables are coupled. We approximate it instead with the following sequential estimate, based on the sequential decomposition above:

$$\begin{aligned} \hat{\boldsymbol{v}} &= \arg\max_{\boldsymbol{v}} P(\boldsymbol{v}|\boldsymbol{I}) & \hat{\boldsymbol{S}} &= \arg\max_{\boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{I}, \hat{\boldsymbol{v}}) \\ \hat{\boldsymbol{R}} &= \arg\max_{\boldsymbol{R}} P(\boldsymbol{R}|\boldsymbol{I}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{S}}) & \hat{\boldsymbol{\ell}} &= \arg\max_{\boldsymbol{\ell}} P(\boldsymbol{\ell}|\boldsymbol{I}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{R}}) \end{aligned} \tag{3}$$

The second challenge is that of actually computing the $\arg\max$ operations in Equation 3. Rather than attempting to model the probability distributions explicitly and maximizing them, we will, instead, model the estimators in Equation 3 as parametric functions:

$$\begin{aligned} \hat{\boldsymbol{v}} &= f_v(\boldsymbol{I}; \theta_v) & \hat{\boldsymbol{S}} &= f_s(\boldsymbol{I}, \hat{v}; \theta_s) \\ \hat{\boldsymbol{R}} &= f_r(\boldsymbol{I}, \hat{v}, \hat{\boldsymbol{S}}; \theta_r) & \hat{\boldsymbol{\ell}} &= f_\ell(\boldsymbol{I}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{R}}; \theta_{\boldsymbol{\ell}}) \end{aligned} \tag{4}$$

The problem of *learning* to estimate the 3D facial parameters thus effectively reduces to that of estimating the parameters θ_v, θ_s, θ_r and θ_l.

Using the common approach, we formulate the learning process for these parameters through an autoencoder. $f_v()$, $f_s()$, $f_r()$ and $f_\ell()$ are, together, viewed as the learnable encoder in the autoencoder, which estimate $\boldsymbol{v}$, $\boldsymbol{S}$, $\boldsymbol{R}$ and $\boldsymbol{\ell}$ respectively. The decoder is a deterministic differentiable *renderer* $\mathcal{R}()$ with no learnable parameters, which reconstructs the original input $\boldsymbol{I}$ from the values derived by the encoder as $\hat{\boldsymbol{I}} = \mathcal{R}(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell})$. The parameters of the encoder are learned to minimize the error between $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}$.

3.2. Facial Parameters Inference

Viewpoint. We first predict the viewpoint parameters from the given image, using a function $f_v(\boldsymbol{I}; \boldsymbol{\theta}_v) : \boldsymbol{I} \rightarrow \boldsymbol{v} \in \mathbb{R}^7$. Here $\boldsymbol{v}$ is used to parameterize the weak perspective transformation [36], including 3D spatial rotation (SO(3)), the translation (xyz coordinates), and the scaling factor.

Shape. The prediction of shape is conditioned on the given image $\boldsymbol{I}$ and the predicted $\boldsymbol{v}$. Since the same face captured with different viewpoints should be correspond to the same facial shape, it is beneficial to exclude as much viewpoint information from the image $\boldsymbol{I}$ as possible before the shape prediction. With the predicted $\boldsymbol{v}$, we can align the image to its canonical view in 2D plane, as shown in Fig. 2 and Appendix A.1. The cropped image is denoted by $\boldsymbol{I} \circ \boldsymbol{v}$. A function $f_s(\boldsymbol{I} \circ \boldsymbol{v}; \boldsymbol{\theta}_s) : \boldsymbol{I} \circ \boldsymbol{v} \rightarrow \boldsymbol{\alpha} \in \mathbb{R}^{228 \times 1}$ with learnable parameter $\boldsymbol{\theta}_s$ is constructed to predict the shape coefficients $\boldsymbol{\alpha}$. The shape coefficients $\boldsymbol{\alpha}$ are defined by a statistical model of 3D facial shape:

$$\vec{\boldsymbol{S}} = \bar{\boldsymbol{S}} + \boldsymbol{U}\boldsymbol{\alpha}, \tag{5}$$

where $\vec{\boldsymbol{S}} \in \mathbb{R}^{3K \times 1}$ is the vectorized $\boldsymbol{S}$, and $\bar{\boldsymbol{S}} \in \mathbb{R}^{3K \times 1}$ is the mean shape. $\boldsymbol{U} \in \mathbb{R}^{3K \times 228}$ is the PCA basis from Basel Face Model (BFM) [27] and 3DFFA [49] for identity and expression variation, respectively. $\bar{\boldsymbol{S}}$ and $\boldsymbol{U}$ are fixed during the training and testing of CEST. With the predicted $\boldsymbol{\alpha}$, the shape $\boldsymbol{S}$ can be obtained using equation 5.

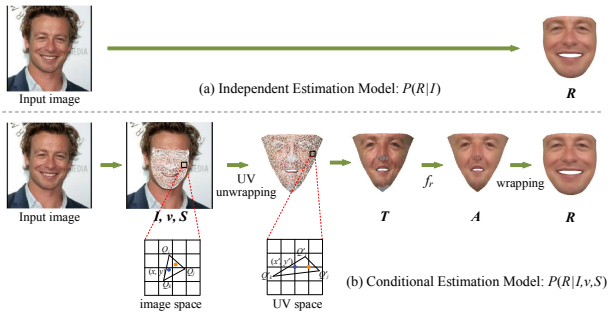


Figure 3: Illustration of generating the UV map of the illuminated texture.

Reflectance. Previous approaches usually predict the reflectance coefficients in a predefined model [39, 38], unwrapped UV map of reflectance [41, 42, 22, 13], or graph representation of the reflectance [44, 48] from the image directly. In CEST, we adopt the UV map representation for reflectance. However, the prediction of the reflectance is conditioned not only on the given image $\boldsymbol{I}$, but also on the predicted viewpoint $\boldsymbol{v}$ and shape $\boldsymbol{S}$.

The process is illustrated in Fig. 2. We first compute the image-coordinate facial shape $\boldsymbol{Q} \in \mathbb{R}^{K \times 2}$ by projecting the world-coordinate facial shape $\boldsymbol{S}$ with viewpoint $\boldsymbol{v}$ using weak perspective transformation. The details of the transformation are given in Appendix A.2, since it is a standard formulation, and not a contribution of this paper. Next, we construct an intermediate representation, *i.e.* UV map of the illuminated texture $\boldsymbol{T}$ [36], which is obtained by unwrapping the given image $\boldsymbol{I}$ based on the predicted face shape $\boldsymbol{Q}$. Subsequently, the UV map of reflectance $\boldsymbol{A}$ is predicted from the illuminated texture $\boldsymbol{T}$ by a reflectance function $f_r(\boldsymbol{T}; \boldsymbol{\theta}_r)$. The reflectance $\boldsymbol{R}$ can be recovered from $\boldsymbol{A}$ by UV wrapping.

The basic idea for computing the $\boldsymbol{T}$ is illustrated in Fig. 3. For each $\boldsymbol{T}(x', y')$ (the pixel values at position (x', y')), we trace its corresponding position (x, y) in $\boldsymbol{I}$. The illuminated texture can be simply obtained by $\boldsymbol{T}(x', y') = \boldsymbol{I}(x, y)$, where bilinear interpolation is used for inferring the pixel values of $\boldsymbol{I}$ at position (x, y) if x or y is not an integer. The computation of (x, y) is as follows. First, the canonical face shape $\bar{\boldsymbol{S}}$ is mapped to the UV space by cylinder unwrapping. We determine the triangle enclosing the point (x', y') on a grid based on the vertex connectivity, which is provided by the 3DMM. The triangle is represented by its three vertices $\boldsymbol{Q}'(i)$, $\boldsymbol{Q}'(j)$, and $\boldsymbol{Q}'(k)$. Since the topology of the facial shape in image space and UV space are the same, the vertices in these two space have one-to-one correspondence. We could easily get the corresponding vertices $\boldsymbol{Q}(i)$, $\boldsymbol{Q}(j)$, and $\boldsymbol{Q}(k)$. Now the position (x, y) can be computed by $x = \kappa_1 \boldsymbol{Q}(i, 1) + \kappa_2 \boldsymbol{Q}(j, 1) + \kappa_3 \boldsymbol{Q}(k, 1)$ and $y = \kappa_1 \boldsymbol{Q}(i, 2) + \kappa_2 \boldsymbol{Q}(j, 2) + \kappa_3 \boldsymbol{Q}(k, 2)$, where the κs

are the coefficients computed by $\boldsymbol{Q}'(i)$, $\boldsymbol{Q}'(j)$, $\boldsymbol{Q}'(k)$, and (x', y') in barycentric coordinate system [6]. The computation details are included in Appendix A.3. For the invisible triangles (caused by self-occlusion), we simply ignore them.

With the illuminated texture $\boldsymbol{T}$, the UV map of the reflectance $\boldsymbol{A}$ can be produced by a function $f_r(\boldsymbol{T}; \boldsymbol{\theta}_r)$, where $\boldsymbol{\theta}_r$ is the learnable parameters. It is worth noting that the input ($\boldsymbol{T}$) and output ($\boldsymbol{A}$) of f_r are spatially aligned in UV space, so the learning process can be greatly facilitated. Subsequently, the reflectance $\boldsymbol{R}$ is obtained by a wrapping function $\boldsymbol{R} = \Psi(\boldsymbol{A})$ [36], which has no learnable parameters, as shown in A.4.

Illumination. Following the previous studies [14, 42], we assume the distant smooth illumination and purely *Lambertian* surface properties [2]. Spherical Harmonics (SH) [28] are employed to approximate the incident radiance at a surface. We use 3 SH bands, leading to 9 SH coefficients. The illumination function is defined as $f_{\boldsymbol{\ell}}(\boldsymbol{I}, \boldsymbol{T}, \boldsymbol{A}; \boldsymbol{\theta}_{\boldsymbol{\ell}}) : (\boldsymbol{I}, \boldsymbol{T}, \boldsymbol{A}) \rightarrow \boldsymbol{\ell} \in \mathbb{R}^{9\times 1}$, which takes the given image, illuminated texture map and UV map of reflectance as input, and produces the illumination parameters.

So far, the 3D face model parameters $\boldsymbol{R}$, $\boldsymbol{S}$, $\boldsymbol{v}$, and $\boldsymbol{\ell}$ are predicted, and we are able to recombine them and render the image by the expert-designed rendering module, *i.e.* $\hat{\boldsymbol{I}} = \mathcal{R}(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell})$.

3.3. Objectives for Self-Supervised Learning

The functions f_s, f_r, f_v, and f_ℓ are modelled by convolutional neural networks (CNNs) with learnable parameters $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_r$, $\boldsymbol{\theta}_v$, and $\boldsymbol{\theta}_\ell$, respectively. Since all the learning modules and expert-designed renderer are differentiable, the proposed framework is end-to-end trainable. The learning objective is to minimize the differences between the original image $\boldsymbol{I}$ and the rendered image $\hat{\boldsymbol{I}}$. Following the practices in previous work, the learning objective does not include the pixels in nonface region, *e.g.* hair, sunglasses, scarf, etc. We identify if a pixel belongs to face or nonface region by a face segmentation network f_{seg}, which is trained on CelebAMask-HQ dataset [23] with the segmentation labels provided in the dataset. Once trained, f_{seg} is fixed during the training and testing of CEST. We denote the effective face region as a mask $\boldsymbol{M}$, so the pixel at position (x, y) is included in reconstruction if $\boldsymbol{M}(i, j) = 1$, and excluded if $\boldsymbol{M}(i, j) = 0$. The photometric loss can be written as

$$\begin{aligned} \mathcal{L}_{ph} &= \mathcal{E}(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}, \boldsymbol{M}) \\ &= \|\boldsymbol{M} \otimes \boldsymbol{I} - \boldsymbol{M} \otimes \hat{\boldsymbol{I}}\|_1 \\ &= \|\boldsymbol{M} \otimes \boldsymbol{I} - \boldsymbol{M} \otimes \mathcal{R}(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell})\|_1, \end{aligned} \tag{6}$$

where $\|\cdot\|_1$ measure the ℓ_1 distance and $\otimes$ denotes the element-wise multiplication. However, if we simply optimize $\mathcal{L}_{ph}$, CEST will learn a degraded solution, where the reflectance $\boldsymbol{A}$ simply copies the pixel values from $\boldsymbol{T}$, and $\boldsymbol{\ell}$ yields an isotropic radiator, radiating the same intensity of radiation in all directions. In this case, CEST does not learn semantically disentangled facial parameters, but leads to a perfect reconstruction for $\hat{\boldsymbol{I}}$.

To avoid this, we adopt the symmetry and consistency constraints for reflectance. The facial reflectance is assumed to be horizontally symmetric and consistent in a video clip. Suppose $\boldsymbol{I}_i$ and $\boldsymbol{I}_j$ are two face images from the same video clip. One of the possible solutions is to add the regularization terms $\|\boldsymbol{R}_i - \boldsymbol{R}_i^{\bowtie}\|$, $\|\boldsymbol{R}_j - \boldsymbol{R}_j^{\bowtie}\|$, and $\|\boldsymbol{R}_i - \boldsymbol{R}_j\|$ to the learning objective, where $\boldsymbol{R}_i^{\bowtie}$ and $\boldsymbol{R}_j^{\bowtie}$ are the horizontally flipped versions of $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$. However, it is difficult to tune loss weights to balance the reconstruction and regularization terms. Instead, we adopt an alternative solution by constructing additional reconstruction terms as constraints [45]. The learning objective for reconstructing $\boldsymbol{I}_i$ and $\boldsymbol{I}_j$ can be written as

$$\begin{aligned} \mathcal{L}_{ph} = &\mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_j, \boldsymbol{S}_j, \boldsymbol{R}_j, \boldsymbol{v}_j, \boldsymbol{\ell}_j, \boldsymbol{M}_j) \\ &+ \mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_j, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_j, \boldsymbol{S}_j, \boldsymbol{R}_i, \boldsymbol{v}_j, \boldsymbol{\ell}_j, \boldsymbol{M}_j) \\ &+ \mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_i^{\bowtie}, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_j, \boldsymbol{S}_j, \boldsymbol{R}_j^{\bowtie}, \boldsymbol{v}_j, \boldsymbol{\ell}_j, \boldsymbol{M}_j) \\ &+ \mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_j^{\bowtie}, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_j, \boldsymbol{S}_j, \boldsymbol{R}_i^{\bowtie}, \boldsymbol{v}_j, \boldsymbol{\ell}_j, \boldsymbol{M}_j) \end{aligned} \tag{7}$$

Stochastic optimization. As can be seen, the number of reconstruction terms is increased dramatically. From n frames of the same video, $2n^2$ reconstruction terms can be constructed. This is not scalable. To address this problem, we propose to optimize the learning objective in a stochastic way. For each training iteration, only a subset of the reconstruction terms are optimized. Specifically, a set of video frames $\{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_N\}$ are randomly sampled from different videos. The frames are grouped by videos, labeled as $\boldsymbol{\xi} = \{\xi_1, \xi_2, ..., \xi_N\}$. For any $\boldsymbol{I}_i$, instead of enumerating all the possible reflectances and obtaining numerous reconstruction terms, we randomly select some other frame from the same video, denoted as $\boldsymbol{I}_j$ (under the condition of $\xi_j = \xi_i$), and use $\boldsymbol{R}_j$ and $\boldsymbol{R}_i^{\bowtie}$ to construct two reconstruction terms for $\boldsymbol{I}_i$. With this strategy, the number of reconstruction terms is reduced from $O(n^2)$ to $O(n)$. Formally, the learning objective can be written as

$$\begin{aligned} \mathcal{L}_{ph} = \frac{1}{N} \sum_{i=1, \xi_j=\xi_i}^{N} &(\mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_j, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) \\ &+ \mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_i^{\bowtie}, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i)). \end{aligned} \tag{8}$$

To stabilize the training of CEST, we use 2D key points via $\mathcal{L}_{kp} = \frac{1}{NN_{kp}} \sum_{i=1}^{N} \sum_{j=1}^{N_{kp}} \|\boldsymbol{Q}_i(k_j) - \boldsymbol{q}_i(j)\|_1$ where $\boldsymbol{q}(j)$ is the set of detected 2D key points on image, and k_j is the index of the vertex associating to the 2D key point. We also regularize the energies of shape coefficients with $\mathcal{L}_{rg} = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{\alpha}_i\|_2^2$. An off-the-shelf landmark detector [7] is used to produce $N_{kp} = 68$ key points for a detected face. The total loss consists of the following terms:

$$\mathcal{L} = \mathcal{L}_{ph} + \lambda_1 \mathcal{L}_{kp} + \lambda_2 \mathcal{L}_{rg} \tag{9}$$

where λ_1 and λ_2 are hyperparameters.

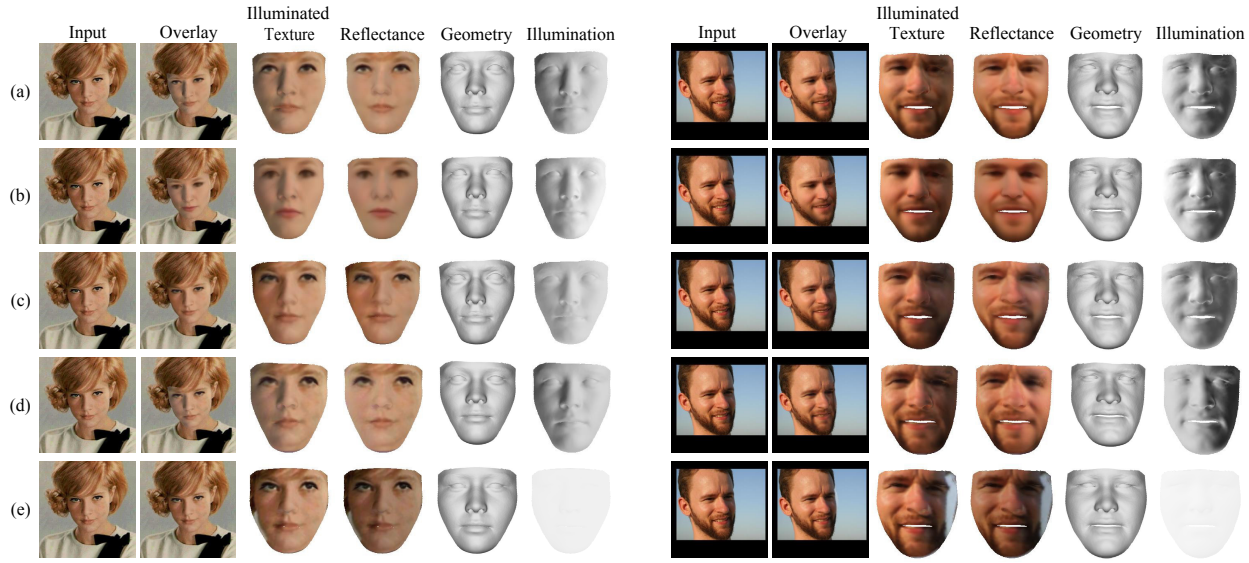


Figure 4: Ablations. (a) CEST with two constraints. (b) Uncoupled CEST with two constraints. (c) CEST with only reflectance consistency constraint. (d) CEST with reflectance symmetry constraint (**the number of video frames is 1**). (e) CEST with no constraint on reflectance. Errata: the row (d) in the ICCV 2021 version mistakenly uses the same results as the row (c). We have fixed it in this version. Please refer to our latest arXiv version for up-to-date results.

4. Experiments

We qualitatively and quantitatively evaluate CEST with ablation experiments and comparisons to state-of-the-art methods [39, 19, 37, 9]. In ablation experiments, we compare CEST to the independent version of CEST (IEST) where facial parameters are estimated in a uncoupled way, and other variants trained with different constraints. Qualitative results include the predicted shape, reflectance, illumination, reconstructed face, etc. we also show the relighted faces, which are obtained by illuminating reflectances with different illuminations. Quantitative results evaluate the qualities of the predicted shape and rendered face. The metrics we used are normalized mean error (NME) [16] and photometric error for shape and rendered face, respectively. NME is defined as the average per-vertex Euclidean distance between the predicted and targeted point clouds normalized by the outer 3D interocular distance. Photometric error is the mean absolute errors between pixel values in the original images and reconstruction images.

4.1. Experimental Settings

For fair comparison, we train two separate CEST models with VoxCeleb1 [26] and 300W-LP [49] respectively. VoxCeleb1 is a video dataset collected from the Internet. The videos of speakers are captured in different in-the-wild scenarios. A subset of 4,727 videos of 267 persons are used in the training, leading to 6,279,609 video frames. The faces in video frames are cropped to the size of 256×256 based on the detected facial key points using [7]. 300W-LP is a synthetic image dataset, containing 122,450 images provided with dense landmarks. Since we focus on self-supervised learning, we only use a sparse set of 68 sparse landmarks as a regularization in training.

Training. The network architectures are given in Appendix B.1. For the training with VoxCeleb1, the minibatch consists of 128 video frames from 32 clips. For each video clip, we randomly selected 4 video frames. The training is completed at 50K iterations. For the training with 300W-LP, the minibatch consisted 128 randomly selected images, and the total iteration is 20K. For both models, we used Adam [21] optimizer with learning rate of 0.001. λ_1 and λ_2 are 1 and 0.1 unless stated otherwise.

4.2. Ablation Experiments

The results of ablation study are shown in Fig. 4. We first present the original and reconstructed image (overlay) for comparison, following by the reflectance, illuminated texture, facial shape (geometry), and illumination in canonical view. More ablations can be found in Appendix B.2.

CEST and IEST. IEST is trained with the same settings as CEST, except the facial parameters are estimated independently from image during training and testing. The results are shown in Fig. 4 (a) and (b), respectively. We can see that CEST produces realistic overlay, disentangled reflectance and illumination, and geometry with personal characteristics and expressions. Compared to CEST, IEST achieves reasonable results, but the reflectances are not as detailed as those from CEST, resulting in inferior overlays and illuminated textures. It validates our hypothesis that the coupled estimation can better formulate the problem and facilitate the learning.

Reflectance symmetry and consistency constraints. We train multiple variants of CEST with only symmetry

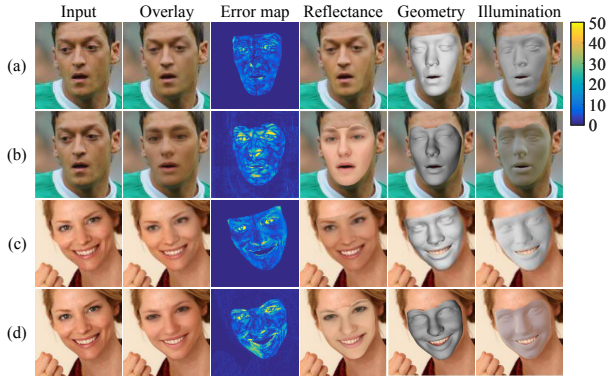


Figure 5: Comparisons with MoFA. (a) and (c) are results from CEST. (b) and (d) are results from MoFA. Images are from CelebA dataset [25]

constraint, only consistency constraints, and without the two constraints, and show their results in Fig. 4 (c), (d), and (e), respectively. Compared (a) and (c) we observe that the reflectance symmetry constraint leads to better reflectance and illumination separation. This is because the horizontally flipped video frames can provide more illumination variations to the training set, enabling CEST to learn to model different illuminations properly. On the other hand, if the reflectance consistency in video clip is not used, the decomposition of reflectance and illumination is not performed well. Some illumination remains around the eyes region in the reflectance (see the right hand side of the Fig. 4 (d)). Lastly, if we do not use any constraints on reflectance, CEST learns the degraded solution (Fig. 4(e)), where the reflectance simply copies the pixel values from the image, and illumination is an isotropic radiator, radiating the same intensity of radiation in all directions. Moreover, we note that the degraded solution also affects the learned facial shape, which has less personal characteristics in Fig. 4 (e).

4.3. Qualitative Results

In this section, we compare CEST to most relevant state-of-art methods with qualitative results. More qualitative results are included in Appendix B.3.

Comparison to MoFA [39]. MoFA is a fully model-based framework. Its representation power is limited by the linear 3DMM model. In addition, all facial parameters from MoFA are independently predicted from the original image. On the contrary, we use a model-free method for reflectance, and the whole inference process is based on coupled estimation. We visualize the overlay, reflectance, geometry, illumination , as well as the errors between input and rendered image (overlays) in Fig. 5. As can be observed, results from MoFA suffer from out-of-subspace reflectance variations. Compared to MoFA, we obtain comparable shape, but significantly better reflectance, illumination, and rendered face by capturing more details.

Comparison to N3DMM [42]. N3DMM generalizes

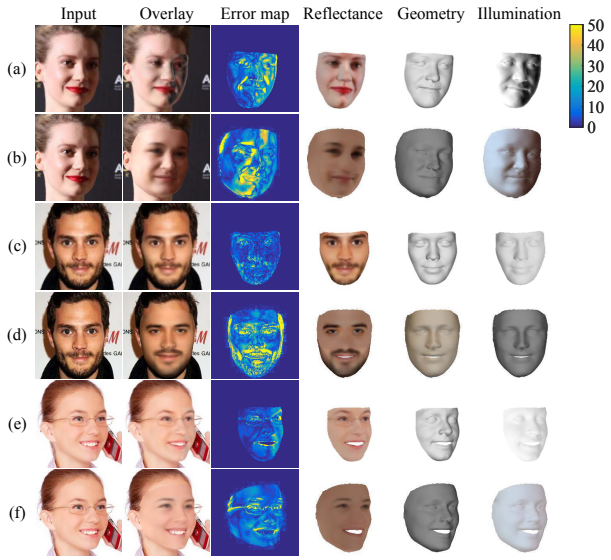


Figure 6: Comparisons with nonlinear 3DMM. (a), (c), and (e) are results from CEST. (b), (d), and (f) are results from N3DMM. Images are from AFLW2000-3D dataset [49]

3DMM model to a nonlinear space and improves the quality of rendered faces. However, N3DMM also infers the reflectance from the input image only, and uses too many heuristic constraints, e.g. reflectance constancy, shape smoothness, supervised pretraining, etc. So their models can only capture low-frequency variations on reflectance. For example, in Fig. 6 (b) the lip stick is missing in the reflectance, and the skin colors in reflectances are almost identical for different persons. These limitations lead to higher reconstruction error. In contrast, our results produce realistic reconstruction, with more accurate reflectance and illumination, as well as lower reconstructed error (Fig. 6).

Comparison to FML [37]. FML properly incorporates video clues in training and can render realistic faces. However, its reconstructed reflectances are prone to an average skin color. In comparison, CEST yields more accurate skin color (see Fig. 7 (a), (c), and (e)) by incorporating the learned shape and viewpoint in the estimation of reflectance. Qualitative results clearly show that our results have more reasonable disentanglement between reflectance and illumination. They also contribute to better visual quality of rendered faces. Notably, there are considerable differences in the eye and nose regions from the overlay in Fig. 7.

Relighting. Since CEST predicts the reflectances of faces, they can be easily re-lighted with different lighting conditions. Fig. 8 shows the re-lit faces in canonical view. In particular, the last two target faces are under harsh lighting, which also examines the illumination removal ability of CEST. The re-lit results again validate that CEST is capable of estimating well-disentangled facial parameters and capturing the reflectance and illumination variations in real-world face images.

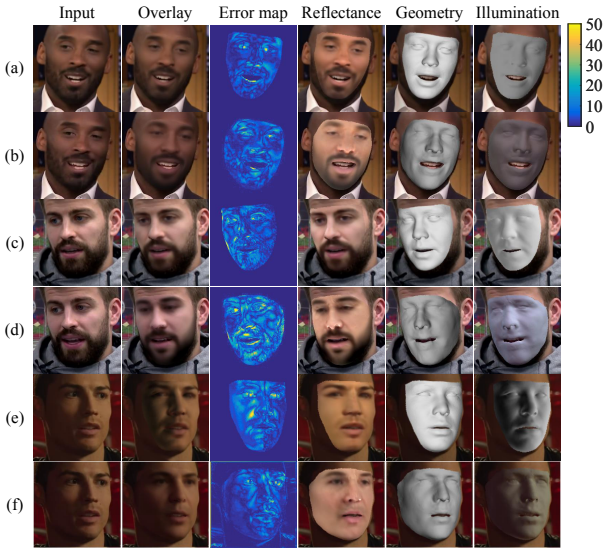


Figure 7: Comparisons with FML. (a), (c), and (e) are results from CEST. (b), (d), and (f) are results from FML. Images are from the video frames in VoxCeleb1 dataset [26]

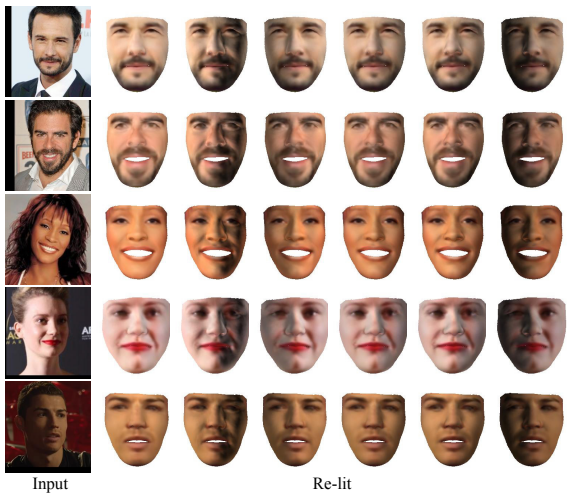


Figure 8: Lighting transfer results.

4.4. Quantitative Results

We first perform quantitative evaluations on the AFLW2000-3D dataset, including 2,000 unconstrained face images with large pose variations. The ground truth of AFLW2000-3D is given by the results from 3DMM fitting, which may be somewhat noisy. The second evaluation is on MICC Florence 3D Face dataset, which consists of high-resolution 3D scans from 53 subjects. We follow the practices in [16] to render 2,550 testing images using the provided 3D scans. Each subject is rendered in 20 difference poses using a pitch of -15, 20 or 25 degrees and a yaw of -80, -40, 0, 40 or 80 degrees.

In order to compare with previous work, NME is computed based on a set of 19,618 vertices defined by [16] in their evaluation. The point correspondences are determined by the iterative closest point (ICP) algorithm [3]. We compute the cumulative errors distribution (CED) curves and compare it to current prevailing methods such as 3DDFA [49], DeFA [24], and PRN [9] on AFLW2000-3D. For MICC, we compare CEST to 3DDFA [49], VRN [16], and PRN [9]. The results are given in Fig. 9. CEST achieves 3.37 and 3.14 NME on AFLW2000-3D and MICC datasets, respectively. More interestingly, our method performs better than the *fully supervised* techniques for shape estimation, e.g. 3DDFA (5.37 on AFLW2000-3D and 6.38 on MICC) and PRN (3.96 on AFLW2000-3D and 3.76 on MICC). Additionally, our method can also estimate facial reflectance and illumination, while both 3DDFA and PRN can not. Compared to N3DMM on MICC dataset, CEST achieves slightly lower NME (3.14 vs. 3.20). Notably, N3DMM uses dense landmarks for supervised pretraining while CEST only uses the 68 sparse landmarks. More quantitative comparisons can be found in Appendix B.5.

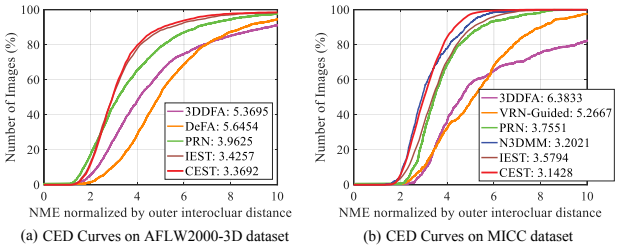


Figure 9: CED curves on AFLW2000-3D and MICC datasets. For example, a point at (4, 63) means 63% of images have NME less than 4.

5. Conclusion and Future Work

We have proposed a conditional estimation framework, called CEST, for 3D face reconstruction from single-view images. CEST addresses the reconstruction problem with a more general formulation, which does not assume conditional independence. We have also proposed a specific decomposition for the conditional probability of different 3D facial parameters. Together with the reflectance symmetry and consistency constraints, CEST can be trained efficiently with video datasets. Both qualitative and quantitative results prove that the conditional estimation is useful. CEST is able to produce high quality and well-disentangled facial parameters for single-view images.

The proposed CEST can be improved from many aspects. Firstly, more accurate and unambiguous facial parameters can be obtained by exploring the temporal information in video. Second, the performance of shape estimation can be boosted by a more advanced morphable model, which also benefits the subsequent estimations of other facial parameters. Moreover, adding perceptual loss could also be an effective way to improve the visual quality of the facial parameters.

References

[1] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80, 2011. 2

[2] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 5

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 8

[4] 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*, pages 187–194, 1999. 1, 3, 11

[5] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5464–5473. IEEE, 2017. 13, 14

[6] O Bottema. On the area of a triangle in barycentric coordinates. *Crux Mathematicorum*, 8(8):228–231, 1982. 5

[7] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 5, 6

[8] 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. 1

[9] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 6, 8

[10] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 1272–1279, 2013. 3

[11] 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. 2, 3

[12] Pablo Garrido, Michael Zollhöfer, Chenglei Wu, Derek Bradley, Patrick Pérez, Thabo Beeler, and Christian Theobalt. Corrective 3d reconstruction of lips from monocular video. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 14, 15

[13] 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 Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019. 4

[14] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 2, 5

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 12

[16] 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. 6, 8

[17] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):394–405, 2010. 2

[18] Ira Kemelmacher-Shlizerman and Steven M Seitz. Face reconstruction in the wild. In *2011 International Conference on Computer Vision*, pages 1746–1753. IEEE, 2011. 2, 3

[19] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *Asian Conference on Computer Vision*, pages 276–292. Springer, 2018. 6

[20] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4625–4634, 2018. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2020. 1, 4

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 5

[24] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1619–1628, 2017. 8

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 7, 15

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 6, 8

[27] 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. 4

[28] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128, 2001. 5

[29] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1268, 2017. 2, 13, 14, 15

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 12

[31] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 15

[32] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017. 13, 15

[33] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018. 1, 2

[34] 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. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 53–70. Springer, 2020. 3, 15

[35] Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 2, 3

[36] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 4, 5

[37] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019. 2, 3, 6, 7, 12, 14, 15

[38] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018. 1, 2, 4, 13

[39] Ayush Tewari, Michael Zollhofer, 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 Workshops*, pages 1274–1283, 2017. 1, 2, 3, 4, 6, 7, 14, 15

[40] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 13, 14

[41] 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. 1, 2, 3, 4

[42] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2, 4, 5, 7

[43] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 14

[44] Huawei Wei, Shuang Liang, and Yichen Wei. 3d dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562*, 2019. 4

[45] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 1, 2, 5

[46] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 1

[47] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013. 1

[48] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1097–1106, 2019. 2, 3, 4

[49] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 2, 4, 6, 7, 8

Appendix

A. Approach

A.1. Image Cropping

The viewpoint $\boldsymbol{v}$ comprises the scale factor $\boldsymbol{v}_1$, 3D spatial rotation parameters $[\boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4]$, and 3D translation parameters $[\boldsymbol{v}_5, \boldsymbol{v}_6, \boldsymbol{v}_7]$. The original image $\boldsymbol{I}$ is cropped to its canonical view in 2D plane with viewpoint v. The cropping is given by $(\boldsymbol{I} \circ \boldsymbol{v})(x', y') = \boldsymbol{I}(x, y)$, where the transformation from (x', y') to (x, y) is formulated in the following.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \exp(\boldsymbol{v}_1) \cdot \cos \boldsymbol{v}_4 & \exp(\boldsymbol{v}_1) \cdot \sin \boldsymbol{v}_4 & \boldsymbol{v}_5 \\ -\exp(\boldsymbol{v}_1) \cdot \sin \boldsymbol{v}_4 & \exp(\boldsymbol{v}_1) \cdot \cos \boldsymbol{v}_4 & \boldsymbol{v}_6 \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} \tag{10}$$

Bilinear interpolation is used if x or y is not an integer.

A.2. Weak Perspective Transformation

The 3D spatial rotation is represented by a rotation vector $\boldsymbol{w} = [\boldsymbol{v}_2; \boldsymbol{v}_3; \boldsymbol{v}_4] \in \mathbb{R}^{3\times1}$: the unit vector $\boldsymbol{u} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$ is the axis of rotation, and the magnitude $\phi = \|\boldsymbol{w}\|_2$ is the rotation angle. The weak perspective transformation is used to project the world-coordinate facial shape $\boldsymbol{S}$ to image-coordinate $\boldsymbol{Q}$, as formulated in

$$\begin{bmatrix} \boldsymbol{Q}(i,1) \\ \boldsymbol{Q}(i,2) \\ \boldsymbol{Q}(i,3) \end{bmatrix} = \exp(\boldsymbol{v}_1) \cdot \left(\boldsymbol{w}\boldsymbol{w}^\intercal \begin{bmatrix} \boldsymbol{S}(i,1) \\ \boldsymbol{S}(i,2) \\ \boldsymbol{S}(i,3) \end{bmatrix} + (\cos\phi) \cdot (1 - \boldsymbol{w}\boldsymbol{w}^\intercal) \begin{bmatrix} \boldsymbol{S}(i,1) \\ \boldsymbol{S}(i,2) \\ \boldsymbol{S}(i,3) \end{bmatrix} + (\sin\phi) \cdot \boldsymbol{w} \times \begin{bmatrix} \boldsymbol{S}(i,1) \\ \boldsymbol{S}(i,2) \\ \boldsymbol{S}(i,3) \end{bmatrix} \right) + \begin{bmatrix} \boldsymbol{v}_5 \\ \boldsymbol{v}_6 \\ \boldsymbol{v}_7 \end{bmatrix}. \tag{11}$$

A.3. Barycentric Coefficients

Given the vertices of a triangle $(\boldsymbol{Q}(i), \boldsymbol{Q}(j), \boldsymbol{Q}(k))$ and its enclosing grid point (x, y) on image. The barycentric coefficients can be computed by

$$\begin{gathered} \boldsymbol{d}_i = \begin{bmatrix} \boldsymbol{Q}(j,1) - \boldsymbol{Q}(i,1) \\ \boldsymbol{Q}(j,2) - \boldsymbol{Q}(i,2) \end{bmatrix}, \quad \boldsymbol{d}_j = \begin{bmatrix} \boldsymbol{Q}(k,1) - \boldsymbol{Q}(i,1) \\ \boldsymbol{Q}(k,2) - \boldsymbol{Q}(i,2) \end{bmatrix}, \quad \boldsymbol{d}_k = \begin{bmatrix} x - \boldsymbol{Q}(i,1) \\ y - \boldsymbol{Q}(i,2) \end{bmatrix}, \\ d_{ii} = \boldsymbol{d}_i^\intercal \boldsymbol{d}_i, \quad d_{jj} = \boldsymbol{d}_j^\intercal \boldsymbol{d}_j, \quad d_{ij} = \boldsymbol{d}_i^\intercal \boldsymbol{d}_j, \quad d_{ki} = \boldsymbol{d}_k^\intercal \boldsymbol{d}_i, \quad d_{kj} = \boldsymbol{d}_k^\intercal \boldsymbol{d}_i, \\ \kappa_2 = \frac{d_{jj} d_{ki} - d_{ij} d_{kj}}{d_{ii} d_{jj} - d_{ij} d_{ij}}, \quad \kappa_3 = \frac{d_{ii} d_{kj} - d_{ij} d_{ki}}{d_{ii} d_{jj} - d_{ij} d_{ij}}, \quad \kappa_1 = 1 - \kappa_2 - \kappa_3. \end{gathered} \tag{12}$$

The barycenteric coefficients κ_1, κ_2, and κ_3 are in the range of $[0, 1]$ if the grid point (x, y) is in the triangle.

A.4. Wrapping Function

The wrapping function $\Psi : \boldsymbol{A} \in \mathbb{R}^{256\times256\times3} \rightarrow \boldsymbol{R} \in \mathbb{R}^{K\times3}$ is defined as $\boldsymbol{R}(i) = \boldsymbol{A}(\boldsymbol{U}(i,1), \boldsymbol{U}(i,2))$, where i is the index for the vertices of a 3D face. $\boldsymbol{R}(i)$ and $\boldsymbol{A}(\boldsymbol{U}(i,1), \boldsymbol{U}(i,2))$ are 3-dimensional vectors. $\boldsymbol{U} \in \mathbb{R}^{K\times2}$ is the coordinates of shape in UV space from 3DMM [4]. Again, bilinear interpolation is used if $\boldsymbol{U}(i,1)$ or $\boldsymbol{U}(i,2)$ is not an integer.

B. Experiments

B.1. Network Architecture

We use standard encoder networks for viewpoint, shape and illumination predictions, and a network similar to U-Net [30] for reflectance prediction. The detailed configurations are given in Table 1. Parameter d is 7 for viewpoint network f_v and 9 for illumination network f_ℓ. Conv $3_{/2,1}$ denotes convoluitonal layer with kernel size of 3, where the stride and padding are 2 and 1, respectively. Each convolutional layer is followed by a Batch Normalization (BN) [15] layer and Rectified Linear Units (ReLU). Bilinear interpolation is adopted for the upsampling operation. Specifically, in Table 1, the layers in brackets are residual blocks. In Table 2, we use shortcut to connect the feature maps of encoder and decoder, but different from U-Net, we use addition rather than concatenation to integrate information in the feature maps. For those encoder output shapes in brackets (*e.g.*, "$[128 \times 128 \times 64]$"), the feature map will be added as a shortcut to the decoder feature map (also with the same brackets).

Viewpoint & Illumination Network			Shape Network		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$256 \times 256 \times 3$	Input	-	$256 \times 256 \times 3$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$128 \times 128 \times 32$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$128 \times 128 \times 64$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$64 \times 64 \times 32$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$64 \times 64 \times 64$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$32 \times 32 \times 64$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$32 \times 32 \times 128$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$16 \times 16 \times 64$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$16 \times 16 \times 128$
$\begin{bmatrix}\text{Conv3} \times 3_{/1,1} \\ \text{Conv3} \times 3_{/1,1}\end{bmatrix}$	BN + ReLU BN + ReLU	$16 \times 16 \times 64$ $16 \times 16 \times 64$	$\begin{bmatrix}\text{Conv3} \times 3_{/1,1} \\ \text{Conv3} \times 3_{/1,1}\end{bmatrix}$	BN + ReLU BN + ReLU	$16 \times 16 \times 128$ $16 \times 16 \times 128$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$8 \times 8 \times 128$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$8 \times 8 \times 256$
$\begin{bmatrix}\text{Conv3} \times 3_{/1,1} \\ \text{Conv3} \times 3_{/1,1}\end{bmatrix}$	BN + ReLU BN + ReLU	$8 \times 8 \times 128$ $8 \times 8 \times 128$	$\begin{bmatrix}\text{Conv3} \times 3_{/1,1} \\ \text{Conv3} \times 3_{/1,1}\end{bmatrix}$	BN + ReLU BN + ReLU	$8 \times 8 \times 256$ $8 \times 8 \times 256$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$4 \times 4 \times 128$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$4 \times 4 \times 256$
Conv $4 \times 4_{/2,1}$	-	$1 \times 1 \times d$	Conv $4 \times 4_{/2,1}$	-	$1 \times 1 \times 228$

Table 1: The detailed CNNs architectures of viewpoint, illumination, and shape networks.

Reflectance Network					
U-Net Encoder (↓)			U-Net Decoder (↑)		
Encoder Layer	Act.	Output shape	Decoder Layer	Act.	Output shape
Input	-	$256 \times 256 \times 3$	Output	-	$256 \times 256 \times 3$
-	-	-	Conv $3 \times 3_{/1,1}$	Tanh	$256 \times 256 \times 3$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$256 \times 256 \times 3$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$128 \times 128 \times 64$	Upsample ($2\times$)	-	$256 \times 256 \times 64$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[128 \times 128 \times 64]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[128 \times 128 \times 64]$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$128 \times 128 \times 64$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$64 \times 64 \times 64$	Upsample ($2\times$)	-	$128 \times 128 \times 64$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[64 \times 64 \times 64]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[64 \times 64 \times 64]$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$64 \times 64 \times 64$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$32 \times 32 \times 128$	Upsample ($2\times$)	-	$64 \times 64 \times 128$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[32 \times 32 \times 128]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[32 \times 32 \times 128]$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$32 \times 32 \times 128$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$16 \times 16 \times 128$	Upsample ($2\times$)	-	$32 \times 32 \times 128$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[16 \times 16 \times 128]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[16 \times 16 \times 128]$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$16 \times 16 \times 128$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$8 \times 8 \times 256$	Upsample ($2\times$)	-	$16 \times 16 \times 256$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[8 \times 8 \times 256]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[8 \times 8 \times 256]$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$4 \times 4 \times 256$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$8 \times 8 \times 256$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$4 \times 4 \times 256$	Upsample ($2\times$)	-	$8 \times 8 \times 256$

Table 2: The detailed CNNs architectures of reflectance networks. Note that, the layers in the decoder (from input to output) are listed from bottom to top.

B.2. More Ablation Studies

We perform more ablations for different settings of CEST. We explore the averaged representations, an approach adopted in [37], for reflectance consistency, where the averaged reflectance of a video clip is used to reconstruct the 3D face in each video frame. Here, we fix the size of minibatch, *i.e.* 128, but vary the number of images from each video clip to 2, 4, and 8.

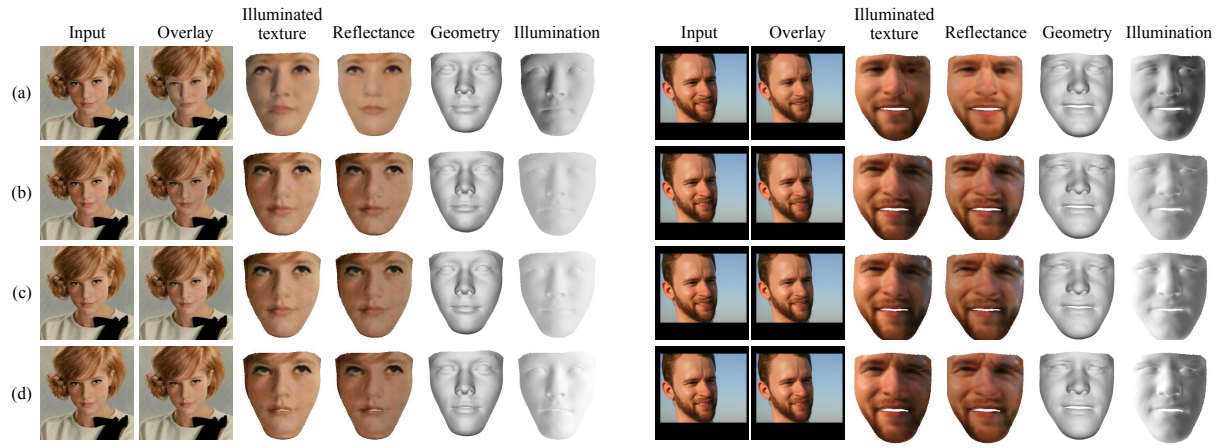


Figure 10: Ablations. (a) CEST with default settings. (b), (c) and (d) Averaged reflectance is used in training and the number of images from each video clips are 2, 4, and 8, respectively. Errata: we mistakenly use the wrong image in the ICCV 2021 version. We have replaced it with the correct one in this version. Please refer to our latest arXiv version for up-to-date results.

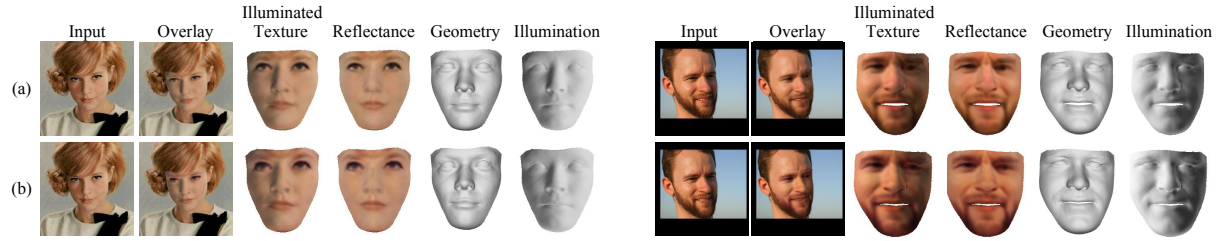


Figure 11: Ablations. (a) CEST with default settings. (b) Reflectance consistency is applied to videos, not video clips.

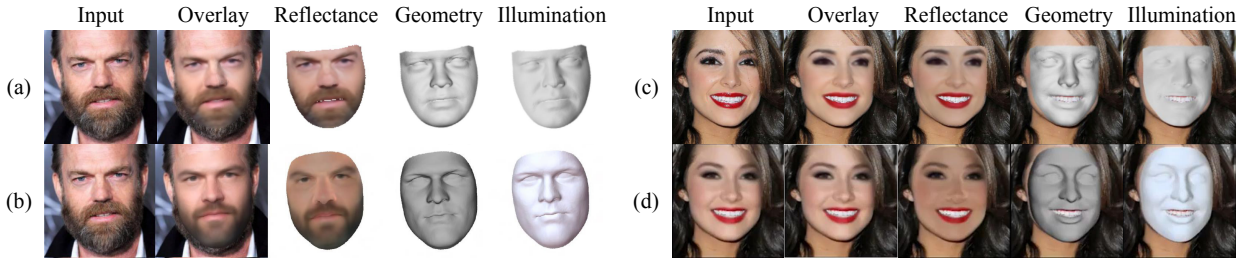


Figure 12: Comparisons to [38]. (a) and (c) Results from CEST. (b) and (d) Results from [38].

Results are shown in Fig. 10 (b), (c), and (d), respectively. As we can see, there are still some illumination in the reflectance, indicating that the averaged representation may not be a good strategy for learning disentangled facial parameters.

Fig. 11 shows the results from CEST trained with reflectance consistency across video. The performance is comparable to those from CEST trained with default setting (reflectance consistency across video clip). It shows that consistency constraint can be generalized to longer videos if the recording environments are not changed dramatically.

B.3. More Qualitative Comparisons

In this section, we show more comparisons to the state-of-art methods [5, 32, 29, 40]. Since there is no publicly available implementations for these methods, we compare to the results presented in their papers.

Overall, CEST produces more stable and reasonable geometries, detailed reflectances, and realistic reconstructions of the 3D faces. As shown in Fig. 12 (a) (b), Fig. 15, Fig. 16, and Fig. 17, the facial shapes predicted by CEST are more accurate in facial expressions and lip closure. In addition, the predicted reflectances show more personal characteristics, but less remaining illumination, as illustrated in Fig. 13 and Fig. 16. Lastly, CEST yields faithful 3D reconstructions, capturing more

details than the other methods (see Fig 14 and Fig 15).

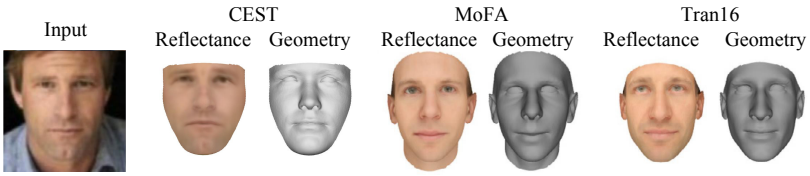


Figure 13: Comparisons to MoFA [39] and [43].

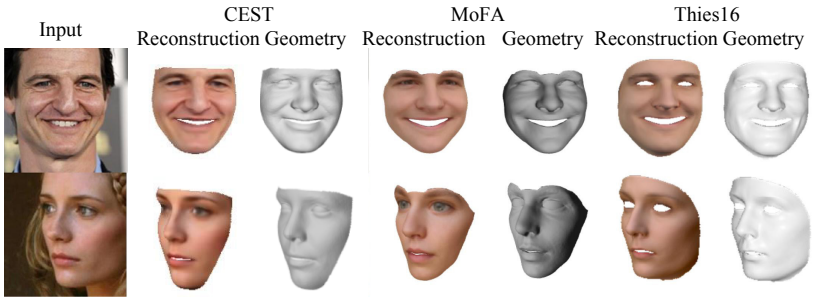


Figure 14: Comparisons to MoFA [39] and [40].

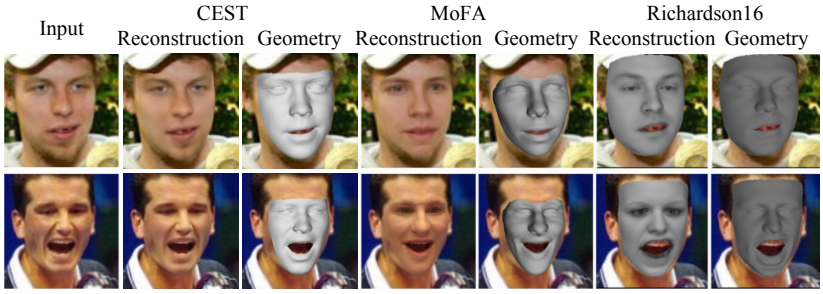


Figure 15: Comparisons to MoFA [39] and [29]. Our estimated shapes show more accurate expressions.

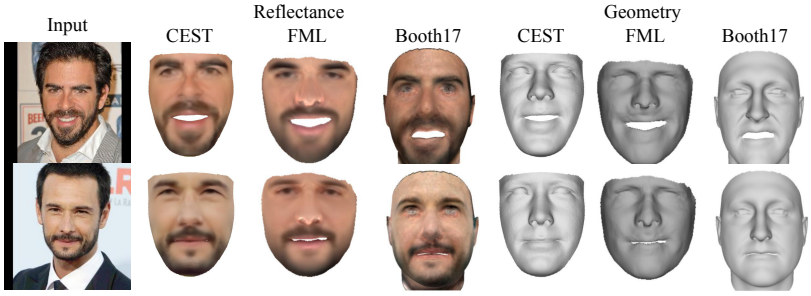


Figure 16: We compare CEST to FML [37] and [5].

B.4. Challenging Cases

We present some examples with dark skin in Fig. 18. Although most people in the training set (VoxCeleb) are Caucasian, CEST still produces reasonable illumination and albedo for these examples. One limitation is that the reconstruction of the non-lambertian surface is inaccurate, e.g. eyes with unusual gaze directions.

B.5. Photometric Error

We compare CEST, IEST, FML [37] and Garrido [12] on overlay face reconstruction. To measure the quality of the overlay images, we compute the average photometric error (R,G,B pixel values are from 0 to 255) between the input face image and

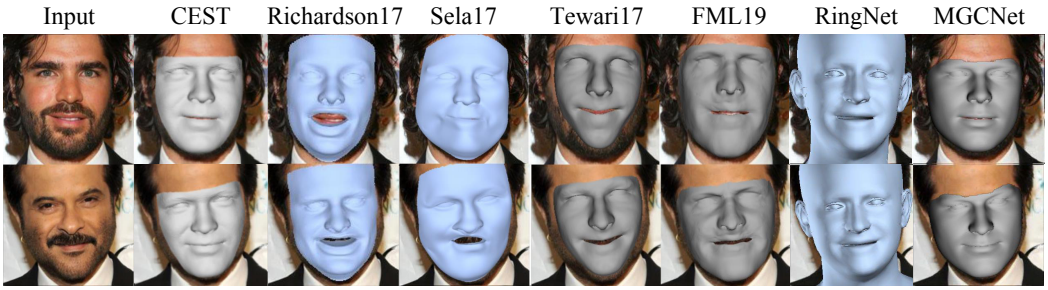


Figure 17: We compare the estimated shapes from CEST to those from [29], [32], [39], [37], [31], and [34] (from left to right). Our estimated shapes are more stable and accurate.

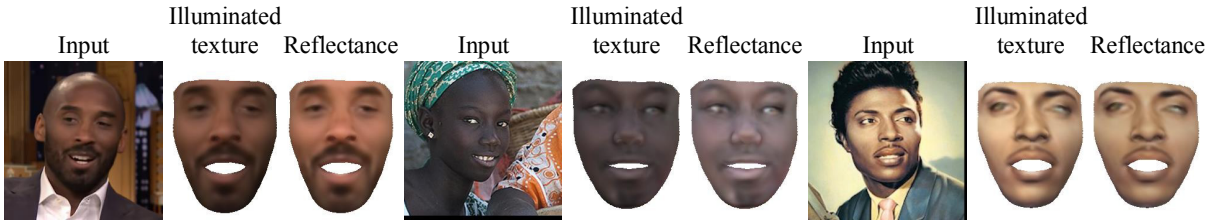


Figure 18: Some challenging examples.

the overlay face image. We experiment on 1,000 images in CelebA dataset [25]. Table 3 shows that the conditional estimation is beneficial for reconstructing the 3D face, and the proposed CEST outperforms existing methods by a large margin.

Method	**CEST**	IEST	FML [37]	Garrido16 [12]
Photometric Error	**10.74**	13.76	20.65	21.95

Table 3: Photometric errors obtained by different methods.