Uncertainty-Aware Mesh Decoder for High Fidelity 3D Face Reconstruction

Gun-Hee Lee[1], Seong-Whan Lee[2]
[1]Department of Computer and Radio Communications Engineering, Korea University, Seoul, South Korea
[2]Department of Artificial Intelligence, Korea University, Seoul, South Korea
{gunhlee, sw.lee}@korea.ac.kr

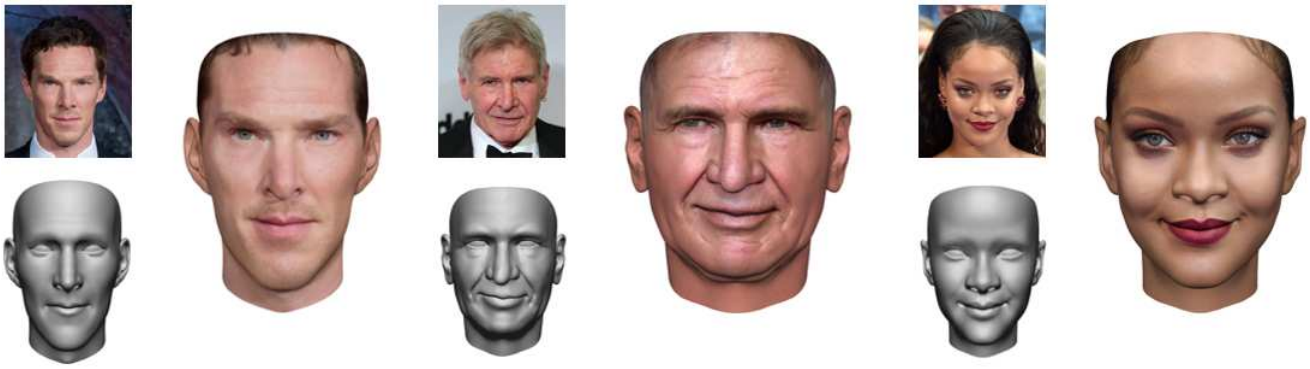

Figure 1: We propose an uncertainty-aware mesh decoder which reconstructs high quality 3D face. Our approach successfully employs Graph CNN and GAN for mesh decoder along with an uncertainty-aware image encoder to reconstruct shape and texture in high fidelity. Our 3D face reconstruction results are best viewed in color.

Abstract

3D Morphable Model (3DMM) is a statistical model of facial shape and texture using a set of linear basis functions. Most of the recent 3D face reconstruction methods aim to embed the 3D morphable basis functions into Deep Convolutional Neural Network (DCNN). However, balancing the requirements of strong regularization for global shape and weak regularization for high level details is still ill-posed. To address this problem, we properly control generality and specificity in terms of regularization by harnessing the power of uncertainty. Additionally, we focus on the concept of nonlinearity and find out that Graph Convolutional Neural Network (Graph CNN) and Generative Adversarial Network (GAN) are effective in reconstructing high quality 3D shapes and textures respectively. In this paper, we propose to employ (i) an uncertainty-aware encoder that presents face features as distributions and (ii) a fully non-linear decoder model combining Graph CNN with GAN. We demonstrate how our method builds excellent high quality results and outperforms previous state-of-the-art methods on 3D face reconstruction tasks for both constrained and in-the-wild images.

1. Introduction

Reconstructing a high quality personalized 3D face can be used for many applications, including face recognition [11, 21, 16, 14, 36], facial motion capture [9] or virtual and augmented reality [34, 33]. However, estimating 3D geometry and texture from a single photograph is still a challenging problem due to the limited 3D scan data and the complex relationship of multiple physical dimensions such as illumination, reflectance, and geometry. To address the above difficulties, Blanz and Vetter [37] introduced 3DMM which relies on additional prior assumption such as constraining faces to lie in a low dimensional subspace. Since then, a lot of 3DMM based methods have been presented and showed impressive results [10, 18, 31, 13, 19, 32]. However, these methods do not generalize well beyond the restricted low dimensional subspace of the underlying linear statistical model.

Recently, many studies have been conducted to bring the concept of DCNN to 3DMMs. This learns model directly from 2D images to better capture in-the-wild variations and increases the representation power. However, these models are still not capable of modeling high fidelity shapes and textures for in-the-wild images. The main reason is that the

3D face reconstruction task still suffers from conflicting requirements between strong regularization for global shape and weak regularization for capturing high level details. We found the fact that we can have special regularization effect by making a change for the embedding step. Unlike recent 3D reconstruction models that use deterministic point representation in the latent feature space, we propose to employ an uncertainty-aware image encoder to inform the decoder what features from the image are uncertain. Let's take an example. When a person describes a particular face from an image, they describe only the facial characteristics of which they are confident. If the eyes are occluded as shown as Figure 2, a person will retain the characteristics of the eyes as uncertain information. As an ideal model would reconstruct confident features with high specificity and uncertain features with high generality, the concept of uncertainty [41, 40] becomes very important for 3D face reconstruction in terms of regularization. These features are also robust to a slight change in the input image (e.g. occlusion, pose or lighting) which could drop the performance for 3D face reconstruction.

Additionally, we focus on a fully nonlinear model and find out that Graph CNN and GAN are effective in reconstructing high quality 3D shapes and textures, respectively. Graph CNN, which directly operates convolutions on non-Euclidean structures such as graphs, manifolds, and meshes, is effective for both obtaining important information from edges and reducing computational complexity. Due to these advantages, it has recently been applied to mesh datasets [30, 29, 23] including 3D face datasets [6, 42]. Meanwhile, as a texture decoder, GAN has recently been proven to be able to represent high fidelity texture and create unobserved views naturally [8]. However, they still have some problems in representing satisfactory results.

In this paper, we propose an uncertainty-aware mesh decoder, which considers the distribution of the input face features and generates a high quality shape and texture using a unified network of Graph CNN and GAN. Along the process, we introduce a novel way to optimize the decoding process using multi-view identity loss and uncertainty-aware perceptual loss, which will further help the model to reconstruct 3D face with high fidelity. The contributions of the paper can be summarized as below:

- We propose to employ an uncertainty-aware image encoder that considers distribution of the face features rather than a deterministic point representation for proper regularization effect.
- We present the unified decoder which combines a detailed shape from Graph CNN with a high-quality texture map from GAN.
- We propose a novel loss function which involves an uncertainty-aware perceptual loss and multi-view identity loss with random projection to further improve the performance of 3D face reconstruction.

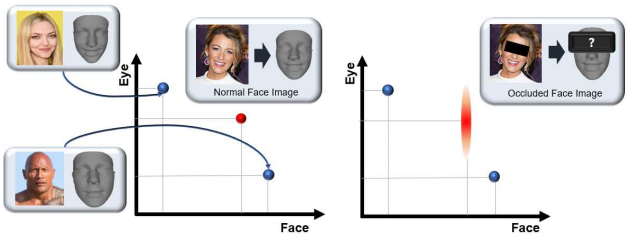

Figure 2: The importance of concerning uncertainty in 3D face reconstruction. Deterministic embeddings represent face as a point estimate without considering uncertain features.

2. Related Work

The history of monocular 3D face reconstruction is quite vast where challenges in building and applying these models are still active research topics. We briefly review the flow of monocular 3D face reconstruction studies from linear to nonlinear approach.

Linear 3DMM. The first concept of the 3DMM was built by Blanz and Vetter [37] that used the Principal Component Analysis (PCA) to represent 3D face shape and texture with linear bases. Since then, there have been many efforts to improve the 3DMM mechanism. Booth *et al.* [18] obtained a richer PCA model with using 10,000 facial scans. Paysan *et al.* [31] improved previous models with using better scanning device and replaced the previous UV space alignment by Nonrigid Iterative Closest Point for registration. Vlasic *et al.* [13] used a multi-linear model of 3D face meshes that separately parameterizes the space of geometric variations due to different attributes. Booth *et al.* [19] used the augmented model with an in-the-wild texture variations. Koppen *et al.* [32] introduced a Gaussian Mixture 3DMM that models the global population as a mixture of Gaussian subspace.

Nonlinear 3DMM. To overcome the aforementioned problem, recent methods focus on nonlinear decoders to go beyond the standard 3DMMs for representation of high quality shape and texture. While both previous optimization-based and learning-based algorithms relied on 3D scan dataset to learn an image-to-parameter or image-to-geometry mapping [4, 17, 27], Tewari *et al.* [7] and Tran and Liu [24] proposed a method of learning 3DMM from 2D images in a self-supervision scheme. As the nonlinear decoder pulled up the reconstruction performance to a satisfying level, very recent studies now try to focus to reconstruct with high details. Tran *et al.* [26] learned additional proxies as means to side-step strong regularization for higher fidelity in 3D reconstruction. Gecer *et al.* [8] used

GAN to reconstruct facial texture with high fidelity. Nevertheless, the method estimates shape with PCA and the ability to represent texture needs to be improved further. Zhou *et al.* [42] first proposed to employ Graph CNN for 3D face reconstruction. However, the method still has some problems in representing satisfactory results.

In this paper, we propose an uncertainty-aware mesh decoder, which considers the distribution of the input face features and generates a high quality shape and texture using a unified network of Graph CNN and GAN.

3. Proposed Method

Our novel 3D face reconstruction approach mainly covers two challenging tasks: (i) balancing the term, generality and specificity by imposing generality for uncertain features and specificity for confident features and (ii) reconstructing high fidelity 3D face with unified decoder of Graph CNN and GAN.

3.1. Uncertainty Encoder

An ideal embedding vector z for neutral shape and albedo should remain consistent for the same identity. However, given the possibility of other either external effect or noises (e.g. pose, blur, occlusion, whitening, illumination) in the input image x, it is not possible to regress the consistent z for all images due to an inevitable shift of the uncertain features. Inspired from the work of Shi *et al.* [41], we propose to employ uncertainty-aware face encoder which can inform the decoder what features from the image are uncertain. This term of uncertainty can also have a special regularization effect which a model is able reconstruct with high fidelity for confident parts and high generality for uncertain parts.

Uncertainty Embedding. In the embedding step, we estimate a Gaussian distribution for $p(z|x_i) = \mathrm{N}\left(\mathrm{z} : \mu_i, \sigma_i^2\right)$ to represent a person's face shape and albedo, where μ_i is the most likely shape feature for the i^{th} input and σ_i is the confidence associated with the corresponding feature. Let us say that we know the μ and the σ for the features of an image pair with same identity (x_i, x_j). Then, we can measure the likelihood of their sharing the same latent vector ($z_i = z_j$, $\Delta z = 0$) which implies that both the shape and the albedo are equal for the same identity:

$$p\left(z_i, z_j\right) = \int p\left(z_i|x_i\right) p\left(z_j|x_j\right) \delta\left(z_i - z_j\right) dz_i dz_j, \quad (1)$$

where z_i and z_j are the latent vector of the shape and the albedo feature of the i^{th} and j^{th} input respectively and the $\delta\left(z_i - z_j\right)$ is a Dirac delta function. We then use log-likelihood which the solution is given:

$$\log p\left(z_i, z_j\right) = -\frac{1}{2}\sum_{l=1}^{D}\left(\frac{\left(\mu_i^{(l)} - \mu_j^{(l)}\right)^2}{\sigma_i^{2(l)} + \sigma_j^{2(l)}} + \log\left(\sigma_i^{2(l)} + \sigma_j^{2(l)}\right)\right) - \frac{D}{2}\log 2\pi, \quad (2)$$

where $\mu_i^{(l)}$ refers to the mean value of the l^{th} feature, $\sigma_i^{2(l)}$ refers to the variance of the l^{th} feature and D denotes the dimensions of feature space. This function can be easily inferred from which the mean value for the difference of two Gaussian's is $\mu_i - \mu_j$ and the variance is $\sigma_i^2 + \sigma_j^2$.

Based on this concept, we use this as a training loss to estimate the uncertainty of shape and albedo. For the training process of estimating this uncertainty value, we use a similar process as [41] proposed. We first fix the value μ, which is earned by pretraining a deterministic 3D reconstruction network without the uncertainty estimation. Then, given a set of images with the same identity, we separately train an additional network that estimates the uncertainty σ. The uncertainty network is a branch network that shares the same input with the bottleneck layer from the encoder. An optimization criterion is used to maximize the above equation for all genuine pair (x_i, x_j), where x_i and x_j is the paired image of equal identity. Formally, we use the loss function to minimize

$$L_{\text{uncertainty}} = \frac{1}{|P|}\sum_{(i,j)\in P} -\log p\left(z_i, z_j\right), \quad (3)$$

where P is the number of all genuine pairs. By this process, the network learns how to estimate the uncertainty σ. We note that this special σ value can act as an attention value for the features and be applied to the loss function.

Expression embedding. For expression embedding, we use a linear blendshape model [12] that combines the facial expression models. We fix this model which is not learned from our work. The 80 blendshape parameters δ are directly applied to the decoder.

3.2. Unified Decoder of Graph CNN and GAN

Mesh Decoder. We propose to employ a Chebyshev spectral Graph CNN [28, 35], which acts directly on the 3D mesh to estimate the face shape by regressing the 3D coordinates of the mesh vertices. It works well with structured graphs with predefined topology. A series of Graph CNN layers operate as the following.

A 3D face mesh can be represented by an undirected and connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \in \mathbb{R}^{n\times 3}$ is a set of n vertices, $\mathcal{E}$ is a set of edges and $W \in \mathbb{R}^{n\times n}$

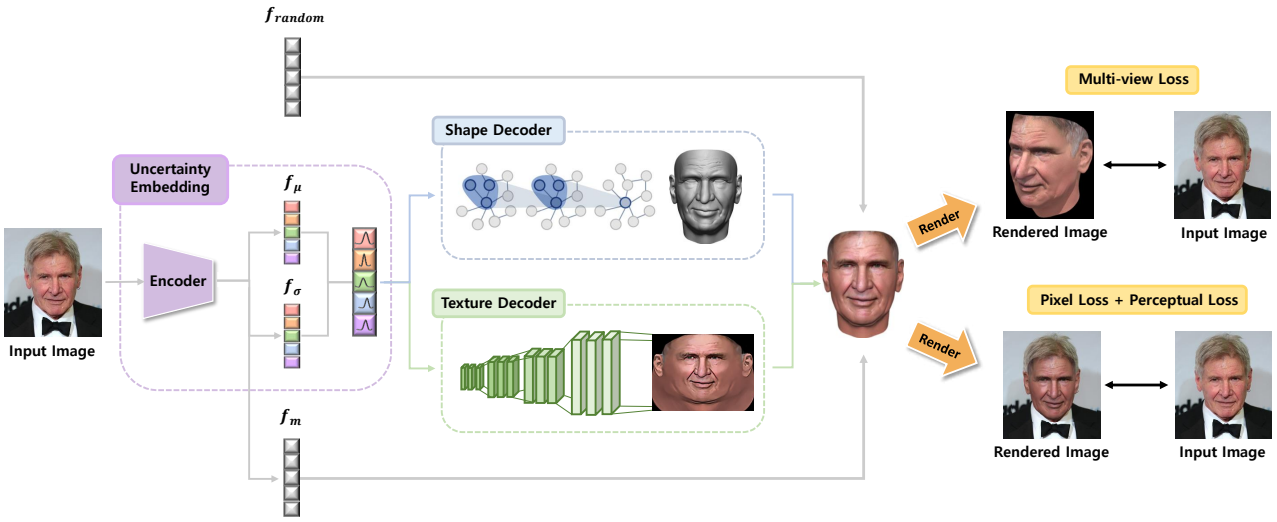

Figure 3: The framework of the proposed method. The model encodes the input image to a projection vector and distribution of the features where the mean represents the most likely feature and the variance represents the confidence associated. The mesh convolution based decoder use this information to reconstruct a 3D face in high-fidelity. We exploit pixel loss, uncertainty-aware perceptual loss, multi-view identity loss, landmark loss, and reconstruction loss.

is an adjacency matrix encoding the connection status between vertices. The normalized graph Laplacian is $L = I_n - D^{-1/2}\mathcal{W}D^{-1/2}$ where $D \in \mathbb{R}^{n\times n}$ is the diagonal matrix with $D_{ii} = \sum_j \mathcal{W}_{ij}$ and I_n is the identity matrix. The Laplacian L is diagonalized by the Fourier bases $U = [u_0, \ldots, u_{n-1}] \in \mathbb{R}^{n\times n}$ such that $L = U\Lambda U^T$ where $\Lambda = \text{diag}([\lambda_0, \ldots, \lambda_{n-1}]) \in \mathbb{R}^{n\times n}$. The graph Fourier transform of $x \in \mathbb{R}^{n\times 3}$ is then defined as $\hat{x} = U^T x$, and its inverse as $x = U\hat{x}$. In Chebyshev Spectral Graph CNN, the graph convolution operation is defined as

$$g_\theta(\Lambda) = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\Lambda}), \tag{4}$$

where $\theta \in \mathbb{R}^K$ is a vector of Chebyshev coefficients and $T_k(\tilde{\Lambda}) \in \mathbb{R}^{n\times n}$ is the Chebyshev polynomial of order k evaluated at a scaled Laplacian $\tilde{\Lambda} = 2\Lambda/\lambda_{max} - I_n$. T_k can be recursively computed by $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0 = 1$ and $T_1 = x$. The spectral convolution can be defined as

$$y_j = \sum_{i=1}^{F_{in}} g_{\theta_{i,j}}(L)x_i, \tag{5}$$

where x_i is the input feature map and y_j is the output feature map.

Based on this concept, we attach the 256-D feature vector extracted by the uncertainty-aware encoder to the 3D coordinates of each vertex in the mean shape. From a high level perspective, the Graph CNN estimates the 3D coordinates of each vertex by using 3D coordinates as the input of each vertex along with the input features.

Texture Decoder. We use a UV map as our texture representation. Each 3D vertex is projected onto the UV space using cylindrical unwrap. In 3D face reconstruction from a single input image, it is important not only to capture a high level of detail, but also to create an unobserved view naturally. In particular, GAN trained with UV map of real textures are shown to be effective in generating realistic UVs while simultaneously generalizing well to unseen data. Gecer *et al.* [8] first proposed to employ GAN for texture decoder. We designed our network using the generator structure suggested by BigGAN [3], which argued that the training of GAN benefits dramatically from large batch sizes. Our network can generate realistic texture map. Figure 4, 5 shows how realistic our texture decoder generates texture maps compared to other methods.

3.3. Differentiable Renderer

To reconstruct a 2D face image from the estimated 3D face, we employ a differentiable renderer [22] based on deferred shading model. Per-vertex attributes such as colors and normal are interpolated at the pixels using the barycentric coordinates and triangle IDs. This approach allows rendering with full perspective and any lighting model.

The 3D textured mesh is projected into a 2D image in the Cartesian coordinates with a camera model. We employed a pinhole camera model that utilizes a perspective transfor-

mation model. The camera parameter can be defined as below:

$$\mathbf{c} = [x_p, y_p, z_p, x_o, y_o, z_o, f]^\top, \tag{6}$$

where $[x_p, y_p, z_p]$, $[x_o, y_o, z_o]$ denote 3D coordinates of camera position, orientation respectively and f is the focal length. Additionally, lighting parameter l is concatenated together with camera parameter as rendering parameter that will be estimated by the uncertainty-aware image encoder. In summary, the rendering parameter $\mathbf{f}_m = \left[\mathbf{c}^T, \mathbf{l}^T\right]^T$, a vector of size 18, is estimated by our uncertainty-aware image encoder and $\mathbf{f}_{random}$ includes randomly taken camera parameter.

3.4. Loss Functions

We propose a novel loss function by combining five terms. It is formulated as below:

$$\begin{aligned}\mathcal{L} = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{unc}\mathcal{L}_{unc} + \lambda_{view}\mathcal{L}_{view} \\ +\lambda_{lan}\mathcal{L}_{lan} + \lambda_{rec}\mathcal{L}_{rec},\end{aligned} \tag{7}$$

where we weight each of loss functions with λ parameters. Our methodology allows us to reconstruct models that are faithful to the input images, depending on the parameter settings, and to reconstruct models that are robust to diverse variations. We optimize all of our parameters so as to minimize our loss function.

Pixel loss. We apply the primitive way of comparing the images in the pixel space with l_1 loss which can be defined as following:

$$\mathcal{L}_{pix} = \left\|\mathbf{I} - \mathbf{I}^{\mathcal{R}}\right\|_1, \tag{8}$$

where $\mathbf{I}$ is the input image and $\mathbf{I}^R$ is the rendered image. This pixel loss L_{pix} enforces the similarity between the input image and the rendered image.

Uncertainty-aware perceptual loss. Simply optimizing for similarity between the images with pixel values can fool the network to produce faces that match closely in the pixel space but look unnatural. The similarity between the identity features from facial recognition network of the input image and the rendered image can help our method to be robust to diverse variations. We use the uncertainty information as weights to compare images in the feature space. The loss gives higher concentration to the confident features that helps reconstruct the corresponding features with higher quality.

$$\mathcal{L}_{unc} = \frac{1}{|D|}\sum_{l}^{D} || \frac{\sigma_i^{2(l)}}{\sum \sigma_i^2}\mu_i^{(l)} - \frac{\sigma_j^{2(l)}}{\sum \sigma_j^2}\mu_j^{(l)} \|_2^2. \tag{9}$$

Multi-view identity loss. Learning to reconstruct 3D face using a single view fits only the observed views. As a result, the reconstructed face looks incorrect when viewed from the different viewpoint with the input image. By adding identity loss with random projection, our method becomes robust to viewpoints. $\hat{\mathbf{I}}^{\mathcal{R}}$ is the rendered image from randomly taken view. $\mathcal{F}(\mathbf{I})$ denotes the uncertainty embedded features of image I.

$$\mathcal{L}_{view} = 1 - \frac{\mathcal{F}(\mathbf{I}) \cdot \mathcal{F}\left(\hat{\mathbf{I}}^{\mathcal{R}}\right)}{\|\mathcal{F}(\mathbf{I})\|_2 \left\|\mathcal{F}\left(\hat{\mathbf{I}}^{\mathcal{R}}\right)\right\|_2}. \tag{10}$$

Landmark loss. We employ a deep face align network [2] $\mathcal{M}(\mathbf{I}) : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{68\times 2}$ to detect landmark locations of the input image and align the rendered mesh onto it by updating the parameters. We apply l_1 loss between the projected and ground truth landmark locations.

$$\mathcal{L}_{lan} = \left\|\mathcal{M}(\mathbf{I}) - \mathcal{M}\left(\mathbf{I}^{\mathcal{R}}\right)\right\|_2. \tag{11}$$

Reconstruction loss. Reconstruction loss is applied to enforcing the shape reconstruction of the mesh. We apply a per vertex l_1 loss between the ground truth shape $\mathbf{S}$ and predicted shape $\hat{\mathbf{S}}$. Empirically we found that using l_1 loss leads to more stable training and better performance than l_2 loss.

$$\mathcal{L}_{rec} = \left\|\mathbf{S} - \hat{\mathbf{S}}\right\|_1. \tag{12}$$

4. Experimental Results

4.1. Datasets

For uncertainty-aware image encoder, we trained our network using the CASIA-Webface dataset [15], which contains 10,575 subjects with over 494k images. For training decoder, we used various datasets including COMA dataset [6], which consists of about 20,000 meshes over 12 different subjects, 300W-LP dataset [38], which contains approximately 60k large pose facial data and the CelebA dataset [43] which is a large scale face attributes dataset with more than 200k celebrity images. We trained texture decoder, Biggan with UV maps from [20].

We perform and conduct qualitative experiments on AFLW2000-3D [38], subset of Casia-Webface dataset, and other various images. We also perform quantitative experiments using the AFLW2000-3D dataset where comparison of the 3D meshes is available for evaluation.

4.2. Ablation Study on Uncertainty

We conduct an experiment to study the effects of uncertainty embedding compared to the most recent 3D face reconstruction methods, which use deterministic embedding.

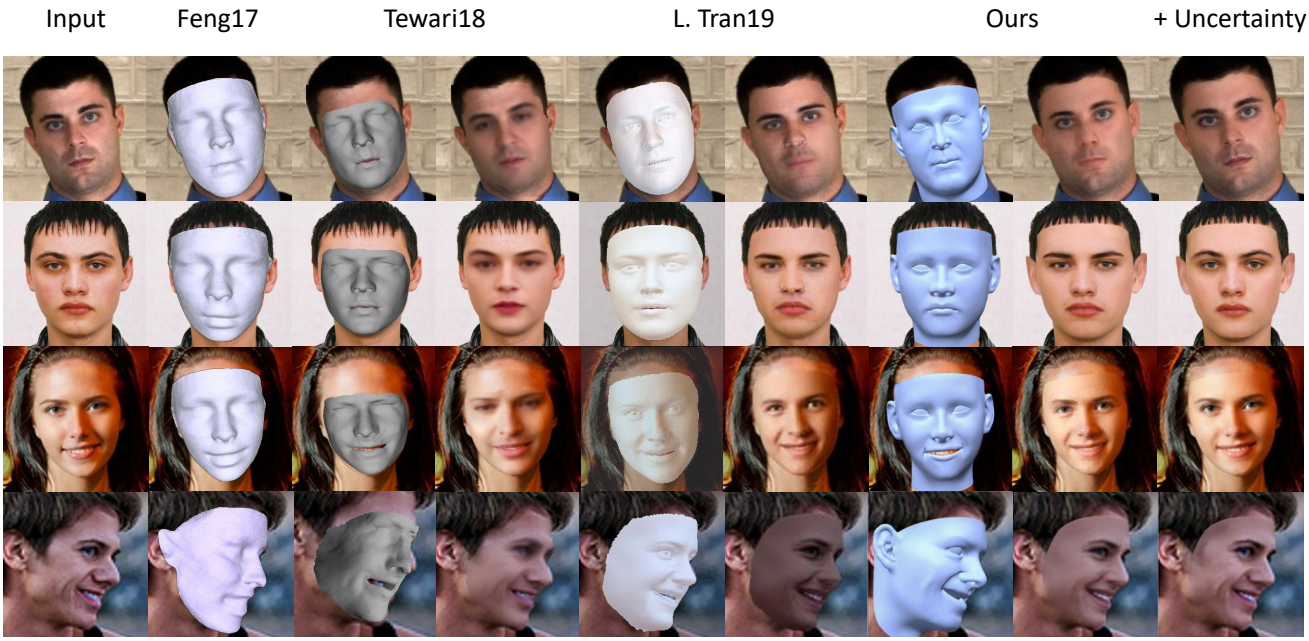

Figure 4: 3D face reconstruction comparison on samples from the AFLW2000-3D dataset [38] and the CelebA dataset [43].

Method	Shape	Texture
Linear [38]	0.0241	0.1287
Nonlinear [25]	0.0146	0.0427
Nonlinear + GL + Proxy [26]	0.0139	0.0363
Ours	0.0129	0.0317

Table 1: Quantitative comparison on shape and texture representation power.

Method	Linear [38]	Nonlinear [25]	PRNet [39]	Ours
NME	5.42	4.12	3.62	2.81

Table 2: Face alignment comparison on the AFLW2000-3D dataset [38].

As shown in Fig. 4, our uncertainty-aware method outperforms recent methods in terms of model's power to represent facial texture. As expected, adding uncertainty to feature embedding has a special regularization effect, which properly controls generality for global shape and specificity for high level details.

4.3. Comparisons to the State-of-the-art

We show our model's power to represent facial shape and texture by comparing the results with other state-of-the-art methods [8, 26, 7, 22, 5, 1] for texture and [39, 26, 7] for shape. As shown in Figure 4, 5, our method outperforms all other methods with high-quality monocular reconstructions of both geometry and texture. We briefly compare the qualitative results of our approach with 8 state-of-the-art methods in 3D face reconstruction task. For linear 3DMM model, the method proposed by A. T. Tran *et al.* [5] estimates the 3DMM parameters using DCNN. However, as the reconstruction subspace is still restricted to the linear bases, the model lacks in representation power for variations in textures. Genova *et al.* [22] trains a regression network from 2D image to 3DMM coordinates using only unlabeled images and synthesized images. Their work proposes three novel loss functions which further helps the reconstruction task in an unsupervised setting. The identity loss for three randomly determined poses gave us the direction for training the identity features which are robust to diverse variations with decreasing the presence of occluded regions of the mesh. However the model is still restricted to the linear subspace which has limited power for representing in-the-wild texture. By contrast our work emphasize the importance of nonlinearity, which further brings the model to go beyond the standard 3DMMs with an ability to represent wide range of shape and texture variation. Jackson et al. [1] tried to avoid using linear 3DMM priors by training a regression network from 2D image to voxel coordinates using an hourglass structure with skip connections. While this strategy had a larger potential for exploring the solution space compared to the linear model, the surface is not smooth and lacks in preserving details. Also volumetric methods discards the relation of meshes which we on the other hand, use Graph CNN to effectively handle face meshes. However, small details of faces disappear as the model estimates a UV position map from an image due to the inevitable smoothing effect of map regression. Feng et

Figure 5: Qualitative comparison with other state-of-the-art 3D face reconstruction methods. The figure shows our model's power to represent facial shape and texture.

al. [39] proposed a straightforward method that simultaneously reconstruct the 3D facial geometry by regressing a 2D representation called UV position map. However, small details of faces disappear as the model estimates a UV position map from an image due to the inevitable smoothing effect of map regression. Again, our work use feature uncertainties along with Graph CNN to have great power to balance regularization on mesh domains which helps to reconstruct shapes with high detail for the confident features.

A. Tewari et al. [7] adopts a self-supervision scheme which breaks out from additional priors, such as statistical face models learned from the limited 3D face data. The 3DMM basis functions are embedded into DCNN and the advantage of 3DMM for regularization is combined with the out-of-space generalization of a learned corrective space. While this model can recover more details than existed 3DMM based methods, the process for model training is attached with strong regularization, which limits their texture representation power for high level details of the face. Our model effectively handles this regularization to create a 3D face model in high-fidelity.

The most related work to our proposed method is Tran

Figure 6: Reconstruction results for CASIA-Webface. The regression network is robust to changes in noise, blur, occlusion and pose.

Figure 7: Reconstruction results for frontal/profile view with multi-view identity loss. Our model is robust in reconstructing all facial areas including the unobserved region.

et al. [26] and Gecer et al. [8]. Tran et al. [26] presents an approach to learn additional proxies as means to avoid strong regularization, as well as, leverages to promote detailed shape and albedo. This method improves the nonlinear 3D morphable model in both learning objective and network architecture which efficiently captures high level details. Our method, however, creates higher level details for both shape and texture by using Graph CNNs which directly operate convolutions on graph based structures and combining it with a GAN model. Also our model use uncertainty information instead of learning additional proxies to loosen regularizations. Gecer et al. [8] harness the power of GAN in order to represent facial texture with high fidelity. GANFIT utilizes GAN to train a very powerful generator of facial texture in UV space. As our work also harness the power of GAN model to create texture maps in high quality, we integrate the textures to the meshes of Graph CNNs and use feature uncertainties to balance the terms for generality and specificity.

4.4. Additional Study on Results

Robust to blur and occlusion. To better understand the power of uncertainty embedding, we additionally conduct an experiment that reconstructs 3D face from ambiguous images. We set the weight of the uncertainty-aware perceptual loss to be large so that our model is robust to diverse variations. As shown in Figure 6, our network yields results that are robust to changes in noise, blur, occlusion, and extreme poses. We show that considering uncertainty brings robustness to varying conditions for a single subject and displays consistent output.

Robust to viewpoint. We study the effects of identity loss according to viewpoint, which measures a distance between the input image and the image rendered with a random projection vector. As shown in Figure 7, images from unobserved viewpoints appear natural. Our method is robust in reconstructing all facial areas including the unobserved region.

5. Conclusion

In this paper, we present an uncertainty-aware mesh decoder which uses uncertainty information to improve 3D face reconstruction task. Our method ensures the decoders to see the uncertain features that can further balance generality and specificity of each features. We also decode shape directly on the mesh domain which is later combined with the generated texture map, where this unified model boost the performance and is suitable to reconstruct both constrained and in the wild images with high details. Our method outperforms previous state-of-the-art methods and this work can be a step toward finding effective feature embedding techniques for 3D face reconstruction.

Acknowledgment

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University), No. 2019-0-01371, Development of brain-inspired AI with human-like intelligence, No. 2014-0-00059, Development of Predictive Visual Intelligence Technology).

References

[1] 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 *ICCV*, 2017. 6

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

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 4

[4] Anh T. Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, 2017. 2

[5] Anh T. Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gerard Medioni. Extreme 3d face reconstruction: Looking past occlusions. In *CVPR*, 2018. 6

[6] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3d faces using convolutional mesh autoencoders. In *ECCV*, 2018. 2, 5

[7] Ayush Tewari, Michael Zollhofer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018. 2, 6, 7

[8] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019. 2, 4, 6, 8

[9] Bindita Chaudhuri, Noranart Vesdapunt, and Baoyuan Wang. Joint face detection and facial motion retargeting for multiple faces. In *CVPR*, 2019. 1

[10] Bon-Woo Hwang, Volker Blanz, Thomas Vetter, and Seong-Whan Lee. Face reconstruction from a small number of feature points. In *ICPR*, 2000. 1

[11] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3d face recognition with a morphable model. In *FG*, 2008. 1

[12] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. In *IEEE TVCG*, 2014. 3

[13] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. In *ACM Trans. Graph.*, 2005. 1, 2

[14] Dihua Xi, Igor T. Podolak, and Seong-Whan Lee. Facial component extraction and face recognition with support vector machines. In *AFGR*, 2002. 1

[15] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. In *CoRR*, 2014. 5

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1

[17] Hyeongwoo Kim, Michael Zollhofer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *CVPR*, 2018. 2

[18] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016. 1, 2

[19] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models in-the-wild. In *CVPR*, 2017. 1, 2

[20] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *CVPR*, 2018. 5

[21] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *CVPR*, 2018. 1

[22] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, 2018. 4, 6

[23] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019. 2

[24] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR* 2018. 2

[25] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. In *TPAMI*, 2019. 6

[26] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019. 2, 6, 8

[27] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 2017. 2

[28] Michael Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeuriPS*, 2016. 3

[29] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2

[30] Nikos Kolotouros, Georgious Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2

[31] Pascal Paysan, Reinhard Knothe, Brain Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *AVSS*, 2009. 1, 2

[32] Paul Koppen, Zhen-Hua Feng, Josef Kittler, Muhammad Awais, William Christmas, Xiao-Jun Wu, and He-Feng Yin. Gaussian mixture 3d morphable face model. In *Pattern Recognition*, 2017. 1, 2

[33] Roger Blanco i Ribera, Eduard Zell, John P. Lewis, Junyong Noh, and Mario Botsch. Facial retargeting with automatic range of motion alignment. In *ACM Trans. Graph.*, 2017. 1

[34] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. In *ACM Trans. Graph.*, 2013. 1

[35] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 3

[36] Un-Sang Park, Hyun-Cheol Choi, Anil K. Jain, and Seong-Whan Lee. Face tracking and recognition at a distance: A coaxial and concentric ptz camera system. In *IEEE Transactions on Information Forensics and Security*, 2013. 1

[37] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 1, 2

[38] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 5, 6

[39] Yao Feng, Fan Wu, Xiahu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 6, 7

[40] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2

[41] Yichun Shi, Anil K. Jain, and Nathan D. Kalka. Probabilistic face embeddings. In *ICCV*, 2019. 2, 3

[42] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: joint texture & shape convolutional mesh decoders. In *CVPR*, 2019. 2, 3

[43] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5, 6