Learning to Restore 3D Face from In-the-Wild Degraded Images

Zhenyu Zhang\textsuperscript{1}, Yanhao Ge\textsuperscript{1}, Ying Tai\textsuperscript{1}, Xiaoming Huang\textsuperscript{1}, Chengjie Wang\textsuperscript{1,*}, Hao Tang\textsuperscript{2}, Dongjin Huang\textsuperscript{3}, Zhifeng Xie\textsuperscript{3,*}

\textsuperscript{1}Tencent Youtu Lab, Shanghai, China
\textsuperscript{2}CVL, ETH Zurich, Switzerland
\textsuperscript{3}Shanghai Film Academy of Shanghai University

Abstract

In-the-wild 3D face modelling is a challenging problem as the predicted facial geometry and texture suffer from a lack of reliable clues or priors, when the input images are degraded. To address such a problem, in this paper we propose a novel Learning to Restore (L2R) 3D face framework for unsupervised high-quality face reconstruction from low-resolution images. Rather than directly refining 2D image appearance, L2R learns to recover fine-grained 3D details on the proxy against degradation via extracting generative facial priors. Concretely, L2R proposes a novel albedo restoration network to model high-quality 3D facial texture, in which the diverse guidance from the pre-trained Generative Adversarial Networks (GANs) is leveraged to complement the lack of input facial clues. With the finer details of the restored 3D texture, L2R then learns displacement maps from scratch to enhance the significant facial structure and geometry. Both of the procedures are mutually optimized with a novel 3D-aware adversarial loss, which further improves the modelling performance and suppresses the potential uncertainty. Extensive experiments on benchmarks show that L2R outperforms state-of-the-art methods under the condition of low-quality inputs, and obtains superior performances than 2D pre-processed modelling approaches with limited 3D proxy.

1. Introduction

3D human face reconstruction has been rapidly advanced during these two decades with applications including human digitalization, animation, and biometrics. The first groundbreaking effort should be the 3D Morphable Model (3DMM) \cite{blanz2003morphable}, which provides reasonable geometry assumptions for modelling. Based on this, the reconstruction can be achieved through optimization on low-dimensional parameters \cite{46,47}. With the development of deep learning, recent works utilize neural networks to regress 3DMM parameters from 2D images \cite{45,70}. Although 3DMM based approaches are further improved with non-linearity \cite{19,25,53–55,60,69} and multi-view consistency \cite{6,10,52,57,61}, they still suffer from several drawbacks: limited amount of subjects (e.g., BFM \cite{44}) with controlled conditions, difficulties on building skin details, anatomic grounded muscles \cite{16} and large variations of identity \cite{71}.

As a result, efforts are made on non-parametric modelling for potential flexibility, which regress face normal or depth directly from an input image without 3DMM assumption \cite{3,49}. More recent works \cite{42,58} disentangle a face into intrinsic factors and accomplishes canonical reconstruction in an unsupervised manner via render-
ing loss [34]. Although the non-parametric methods capture more detailed and distinct facial structures, they usually suffer from degradation of appearance as facial clues are only provided by input images without 3DMM prior. As shown in Fig. 1, degraded inputs significantly reduce the reconstruction accuracy. In practice, in-the-wild facial images often have low resolution and quality due to unsatisfactory equipment or a low proportion of the whole scene. On top of these, we argue that high-quality face modelling against degraded images is practical and crucial for non-parametric methods. To tackle this problem, a direct way is using pre-trained super-resolution models [37, 40] to process the degraded images. However, these models only tackle 2D appearance but fail to enhance inherent 3D clues. As illustrated in Fig. 1, 2D pre-processing cannot well improve 3D reconstruction accuracy, showing unsatisfactory visual results and fragile performances on pose variation. While 3D texture completion methods [22, 68] inpaint the missing facial region, they cannot enhance the geometry.

In this paper, we propose a novel Learning to Restore (L2R) 3D face framework to improve 3D face modelling against limited image quality. L2R achieves such a goal by mining 2D facial priors from pretrained GANs for the propagation of 3D texture/geometry clues. The framework is conducted in a mutual paradigm to iteratively boost 3D texture and geometry modelling from a simple proxy. Concretely, to constrain the generated texture with suitable content and 3D UV-position, L2R encodes input images and albedo proxy to StyleGAN [33] generator, providing style codes and spatial prior, respectively. In this way, L2R urges StyleGAN to predict diverse clues on modelling realistic 3D albedo beyond degraded input. Further, benefited from the 3D textures, L2R learns high-resolution facial shapes and displacement maps to enhance facial details without predefined topology. As 3D texture and geometry modelling complement each other via rendering, we mutually optimize these two procedures with a novel 3D-aware adversarial loss, which enhances the consistency of prediction. Extensive experiments demonstrate that L2R models superior texture and geometry from low-resolution images than state-of-the-art and 2D pre-processed methods, and obtains competitive results to models without degradation.

In summary, this paper has contributions in followings:

i) A novel Learning to Restore (L2R) 3D face framework is proposed to model high-quality 3D faces from degraded images in an unsupervised manner. In contrast to 2D appearance processing methods, L2R is able to enhance inherent 3D clues on texture and geometry reconstruction.

ii) With a novel albedo restoration network, L2R mines 2D generative facial priors to complement the lack of facial clues and models 3D finer textures.

iii) Based on the restored 3D texture, L2R uses a novel geometry refining network to model detailed facial depth

with a displacement map and enhances the 3D proxy.

2. Related Works

In Table 1, we compare recent 3D face modelling approaches, among which our method tackles a more challenging setting without shape assumption, and models 3D face from degraded input images.

| 3D Face Reconstruction: | As a long-standing problem, 3D face reconstruction is firstly developed by 3DMM [8, 46]. These optimization based methods are further improved by deep neural network [15, 20, 38, 45, 70]. With the differentiable renderer [34], models leverage image reconstruction loss to get rid of ground truth dependency [26, 54]. Recently, 3DMM approaches are further improved with non-linearity [19, 25, 53, 60, 69] and multi-view consistency [14, 52, 57].

To improve accuracy beyond 3DMM, non-parametric methods, e.g., shape-from-shading algorithm [64], is also able to model 3D face without 3DMM assumption. With the success of deep learning, such an algorithm is improved by SFS-Net [49] for modelling intrinsic facial factors. Data-driven methods [3, 30, 63] are also proposed to directly learn face geometry supervised by real or synthetic ground truth. However, they cannot model 3D geometry of full face. More recent works [58, 67] use weakly symmetric constrains to predict canonical intrinsic factors from facial images. GAN2Shape [42] avoids such symmetric constraint but brings heavy per-image optimization. LiftedGAN [51] transforms the framework to a generative model but also needs optimization to address real-world images. While these non-parametric methods cannot tackle low-quality input images, our method improves the 3D modelling under challenging conditions. Moreover, our method is conducted in an end-to-end manner without per-image training.

3D Texture Completion: Facial texture synthesis has been extensively studied [21, 25, 50]. Nevertheless, the performance of these methods are limited due to the ambiguity of monocular images. Therefore, approaches are proposed to inpaint texture from visible facial appearance [12, 24]. Recently, [68] propose a framework for completion by rotating and rendering. [23] propose a novel one-shot completion method using 2D face generator. In summary, these methods require high-quality input, per-image optimization and cannot improve facial geometry. In contrast, our
method tackles low-resolution input, geometry refining and realizes end-to-end inference.

GAN Inversion: Images can be embedded to the latent space of pre-trained StyleGAN [32, 33], and efforts started from [1, 2, 7]. Based on such approach, methods are proposed to restore degraded images [28, 43]. In practice, such inversion models require extra optimization procedure, or lack of spatial constraints only depended on style codes. In contrast, our method needs no extra optimization and provides UV-spatial guidance to texture generation.

3. Learning 3D Proxy

Our L2R is able to improve the quality of a 3D proxy against image degradation. Theoretically, such proxy can be arbitrary non-parametric model. Here we select Unsup3D [58] as 3D proxy, as it requires no supervision, limited constraints and training cost. As discussed in Sec. 1, Unsup3D suffers from noise and blurring from input images, from which it predicts 3D model with unsatisfactory reality and consistency. However, the UV-relationship it provides can be leveraged by L2R to guide high-quality textured and arbitrary non-parametric model. Here we select Unsup3D as 3D proxy, as it requires no supervision, limited constraints and training cost. As discussed in Sec. 1, Unsup3D suffers from noise and blurring from input images, from which it predicts 3D model with unsatisfactory reality and consistency. However, the UV-relationship it provides can be leveraged by L2R to guide high-quality texture generation.

Unsup3D disentangles a facial image \( I \) into intrinsic factors \((d, a, \omega, l)\) including a canonical depth map \(d \in \mathbb{R}^4\), a canonical albedo image \(a \in \mathbb{R}^3\), a global light direction \(l \in \mathbb{S}^2\) and a viewpoint \(\omega \in \mathbb{R}^6\). Each factor is predicted by a separate network which we denote as \(\Phi^d, \Phi^a, \Phi^\omega, \Phi^l\). Then, the 3D face can be reconstructed using these factors by lighting \(\Lambda\) and rasterization \(\Pi\) as follows:

\[
\hat{I} = \Pi(\Lambda(a, d, l, d, \omega)),
\]

(1)

\(\Pi\) is achieved by a differentiable renderer [34]. The learning is performed via image reconstruction loss which encourages \(I \approx \hat{I}\). To represent full frontal face and get canonical albedo/depth, the framework utilizes a weakly symmetric constraint by horizontally flipping:

\[
\hat{I}' = \Pi(\Lambda(a', d', l, d', \omega)),
\]

(2)

where \(a'\) and \(d'\) are the flipped version of \(a, d\). Meanwhile, the objective \(I \approx \hat{I}'\) is also encouraged. As practical faces may be asymmetric, the framework predicts confidence maps \(\sigma, \sigma' \in \mathbb{R}_+\) by \(\Phi^\sigma\) and calibrates the loss as follows:

\[
\mathcal{L}(I, \hat{I}, \sigma) = -\frac{1}{|\Omega|} \sum_{\Omega} \ln \frac{1}{\sqrt{2\sigma}} \exp\left(-\frac{\sqrt{2}(I - \hat{I})}{\sigma}\right),
\]

(3)

where \(\Omega\) is normalization factor. The flipped version \(\mathcal{L}(I', \hat{I}', \sigma')\) is also calculated. In this way, 3D faces are modeled from images in an unsupervised manner without 3DMM assumption. For clearness, we denote \(a_o, a'_o, d_o, d'_o\) as the predicted albedo/depth proxy, and \(l_o, \omega_o\) as the light/viewpoint proxy. The degraded input and high-resolution ground truth is denoted as \(I, I_{gt}\).

4. Learning to Restore 3D Face

In this section, we mainly describe the proposed Learning to Restore 3D Face (L2R) framework. Given a low-resolution image, our aim is to mine 2D generative facial priors for fine-grained 3D texture and geometry modelling. As illustrated in Figs. 2 and 3, the framework has two modules: Albedo Restoration Network (ARN) and Geometry Refining Network (GRN). We further propose a 3D reality loss to improve the 3D consistency of predictions, and introduce a mutual learning strategy for effective optimization.

4.1. Albedo Restoration Network

To tackle the degradation of input images, recent works [9, 28] show that a pre-trained StyleGAN [32, 33] is able to provide complementary priors. However, StyleGAN only contains 2D texture clues and struggles to generate decoupled 3D information. As a result, we propose the ARN which urges StyleGAN to provide 3D canonical facial albedo clues. The ARN is illustrated in Fig. 2, where the style code and spatial guidance respectively guarantee the content of input image and explicit 3D position.

Style Code Injection: To guarantee the suitable characteristics of restored albedo as the input image \(I\), corresponding style code should be extracted to guide the pre-trained StyleGAN. We use a style encoder with the same architecture as those of \(\Phi^d, \Phi^a\) to get high-level features from \(I\), and then use the same fully-connected mapping network as [33] to obtain style code \(c\). Here we predict multiple codes, i.e., for a StyleGAN with \(n\) stages, we generate codes as \(\{c_i\}_{i=1}^n\). Then we inject the code \(c_i\) to AdaIN [29] of the \(i\)-th style conv-layer. In this way, StyleGAN obtains ‘styles’ at each stage, which contributes to generating priors with suitable multi-level attributes. Note that, different from inversion methods [1, 2], we do not require the style codes to strictly recover \(I\), but provide reasonable facial information.

Spatial Guidance Injection: The prediction of ARN should be in canonical view as \(a_o\) to represent 3D full-face texture, thus StyleGAN should be guided to generate priors in UV-space. To achieve this, we extract features from the albedo proxy as spatial guidance. As illustrated in Fig. 2, the albedo encoder has the same architecture as \(\Phi^a\), which extracts multi-scale features that denoted as spatial guidance \(\{g_i\}_{i=1}^k\). We then propagate the spatial guidance to the first \(k\) corresponding stages of StyleGAN. Denote \(f_i\) is the input feature of the \(i\)-th StyleGAN stage, the spatial guidance injection can be formulated as \(f'_i = Conv([g_i, f_i])\), where \(f'_i\) is the output of StyleGAN’s conv-layer. We only inject the guidance to early stages of StyleGAN, as propagating the features of albedo proxy to higher layers may apply too much dependency and limit the quality of generated priors. In this way, StyleGAN obtains UV-relationship to provide canonical 3D facial priors.

Prior Decoder: To further organized the information
provided by StyleGAN layers, we propose a prior decoder to process the output features. As information of higher layers of StyleGAN is more correlated to image appearance, we leverage output features from last \( m \) stages of StyleGAN as facial priors. As illustrated in Fig. 2, we denote the priors as \( \{ p_i \}_{i=1}^{m} \). Then the priors are fed into the prior decoder to generate the final restore albedo \( a_r \). The prior decoder can be formulated as:

\[
h_i = \begin{cases} \text{Conv}(p_i), & i = 1 \\ \text{Conv}([h_{i-1}, p_i]), & \text{otherwise} \end{cases}
\]  

(4)

where \( h_i \) is the output feature of StyleGAN’s conv-layer. After each conv-layer, we upsample \( h_i \) to match the feature size. \( a_r \) is used to re-render high-resolution image that we denote as \( I_r \) with the proxy of \( \omega_o \) and depth \( d_o \) by Eqns. (1), (2). In our method, we recover albedo to a reasonable size which is 8-time larger than original input to avoid huge training and rendering burden. Hence we use a reasonable size which is 8-time larger than original input to analyse our method.

4.2. Geometry Refining Network

As discussed in Sec. 1, 2D appearance enhancing cannot provide 3D texture clues which are directly beneficial to geometry modelling. Hence, we propose Geometry Refining Network (GRN) to leverage recovered 3D texture clues for detailed facial shape modelling, illustrated in Fig. 3. An encoder-decoder network is utilized to predict pure normal \( n_r \) and recovered canonical texture \( t_r \). Here, features from \( t_r \) provide high-quality 3D texture clues in canonical view, while \( I \) and \( n_o \) give priors and content in original domain. In practice, we find that only predicting pure depth/normal cannot well reconstruct facial details but yields over-smooth results. Hence, we learn displacement map to explicitly model the fine details. Such idea is inspired by 3DMM based methods \([11, 18, 59]\) but...

![Figure 2. Overview of the proposed Albedo Restoration Network. Spatial guidance and style codes are extracted from original canonical albedo \( a_o \) and low-resolution image \( I \), respectively, and then fed to a pre-trained StyleGAN. Style code helps StyleGAN to generate priors with similar style to the target, and spatial guidance constrains the UV-relationship of the prediction to generate 3D full-face albedo. Facial priors of higher-level layers of StyleGAN is then fed into the prior decoder to predict restored albedo. The restored image \( I_r \) with similar style to the target, and spatial guidance constrains the UV-relationship of the prediction to generate 3D full-face albedo.](image1)

![Figure 3. Overview of the Geometry Refining Network (GRN). An encoder-decoder network is utilized to predict pure normal \( n_r \) and displacement map \( m \) from input \( I \), normal proxy \( n_o \), and recovered canonical texture \( t_r \).](image2)

4.3. Training

Besides loss function in Eqn. (3), to enhance 3D modelling quality, we propose a cross-view consistency loss. We re-render a restored image using randomly sampled \( \hat{\omega}, \hat{I} \), and denote it as \( \hat{I}_r \). Given the high-resolution ground truth \( I_{gt} \), we use adversarial loss \([4]\) with a same discriminator \( D \) as \([33]\) to compare \( I_r \) and \( I_{gt} \). This objective can be formulated as:

\[
L_c = \min_G \max_D \mathbb{E}[\log(D(I_r))] + \mathbb{E}[\log(1 - D(G(\hat{I}_r)))]
\]  

(6)

where \( G \) is our framework as generator. To suppress artifact in predicted displacement map, we use a Laplacian smooth-
ness regularity which can be denoted as $\nabla |m|$. The final loss is then formulated as:

$$L_{afl} = \mathcal{L}(I, I_{gt}, \sigma) + \mathcal{L}(\hat{I}, I_{gt}, \sigma') + \alpha L_c + \beta \nabla |m|, \quad (7)$$

where $\alpha, \beta$ are weighted constants.

With the proposed ARN and GRN, we recover detailed 3D texture and geometry from degraded images. Actually, the modelling procedures of ARN and GRN are complementary with each other: better texture provides reliable clues for geometry reconstruction, while finer shape details contribute to appearance prediction via rendering procedure. As a result, we propose a mutual learning strategy to boost the optimization. We first optimize ARN, and then train GRN with ARN freezed. During this stage, we directly use $\omega_o, l_o$ proxy predicted from $\Phi^{\omega}, \Phi^l$. Then we jointly optimize ARN, GRN and $\Phi^{\omega}, \Phi^l$ to mutually improve the performance of texture and geometry modelling. With such strategy, the ambiguity of intrinsic learning is reduced, and detailed clues of StyleGAN are successfully propagated to the whole modelling procedure.

5. Experiment

Dataset. Following Unsup3d [58], the model of 3D proxy is pretrained on CelebA [39] which has 200k in-the-wild human face images. We select 160k/20k/20k images as train/val/test set, respectively. Once the 3D proxy is prepared, we train L2R on CelebAMask-HQ [36] and FFHQ dataset [32] which contains 30k and 70k images, respectively. For FFHQ, we randomly select 30k images to reduce the training time burden. The final dataset is combined with 60k images, where we select 40k/10k/10k as train/val/test. The images are resized to 256×256 as high-quality ground truth $I_{gt}$, and are further resized to 32×32 with different blurring and smoothing as degraded input $I$.

For evaluation on facial geometry, we perform testing on 3DFAW [27, 31, 65, 66], BFM [44] and Photoface [62] dataset. 3DFAW contains 23k images with 66 3D keypoint annotations, and we use the same protocol as [58] to calculate depth correlation metric for testing. For BFM dataset, we use the same generated data released by [58] to evaluate predicted depth maps. Photoface dataset contains 12k images of 453 people with face/normal image pairs, and we follow the protocol of [3, 49] to evaluate the quality of modeled facial normal.

Implementation Details. We use the same architecture of $\Phi^{\omega}, \Phi^l$ as [58] to predict pose and light. For learning confidence map at output size of 256×256, we use a similar network $\Phi^{\sigma}$ but with extra upsampling-conv operations. We use StyleGAN2 [33] which is officially pretrained on FFHQ of 1024×1024 in ARN to provide generative facial priors. This model contains 9 stages, thus we set $n = 9$ in style code injection of Fig. 2. The spatial guidance contains features from 4×4 to 32×32 extracted from initial albedo, thus we use $k = 4$ in $\{g_i\}_{i=1}^k$. The priors are features from 64×64 to 1024×1024, thus we set $m = 4$ in $\{p_i\}_{i=1}^m$. For the priors of 512×512 and 1024×1024, we downsample them to 256×256 to match the output size. For final loss in Eqn. (7), we set $\alpha, \beta = 0.1$. To optimize L2R framework, we first train ARN for 30 epochs, then freeze ARN and train GRN for another 30 epochs. The final mutual learning of ARN and GRN is performed for 20 epochs. The learning is conducted by Adam solver [35] with learning rate of $1e−4$ and batch size of 16 on one NVIDIA Tesla V100 GPU. More details can be found in the appendix.

Evaluation Protocol. For predicted facial geometry, following [58, 67], we use Scale-Invariant Depth Error (SIDE) [17] and Mean Angle Deviation (MAD) for evaluating depth and normal. For evaluation on modeled texture, we calculate Structural Similarity Index (SSIM) [56] on facial regions between ground truth images and rendered ones. To evaluate cross-view consistency of the reconstruction, we render images of original and frontal pose, ±45° of yaw and pitch angles, and calculate mean cosine-similarity of their encoded representations of Arcface [13] between that of original high-quality image. To fairly assess the cosine similarity, we use a same cropping protocol as Arcface to process the images. All the evaluation is performed on scale of 256×256.

5.1. Ablation Study

Geometry: In this section we first analyse how the proposed methods influence the geometry modelling. The results are illustrated in Table 2. Our full model obtains best performance. Lines (2)-(6) and (7)-(9) reveal the influence of ARN and GRN, respectively. In line (2), we observe that even with a same architecture, the lack of pre-trained StyleGAN prior significantly increases the errors. Such a phenomenon also demonstrates that L2R successfully leverages facial clues contained in StyleGAN to boost geometry modelling. In line (3), we remove the StyleGAN but apply DFDNet [37] to pre-process the low-resolution images as input. This approach brings lower errors, but still worse than our full methods. This experiment indicates that only pre-processing 2D appearance cannot tackle the 3D reconstruction from degraded images, while StyleGAN provides reliable facial priors to compensate for the lack of clues. Lines (4)-(6) reveal the proposed components in ARN also contribute to better performance. In line (7), we observe that the restored image plays an important role, which provides complementary details for finer reconstruction. Furthermore, Line (8) reveals displacement map increases the accuracy beyond pure normal, while Line (9) demonstrates the effectiveness of our cross-view loss in Eqn. (7). We further make visual comparisons in Fig. 4. The input images are blurred, and the 3D proxy suffers from such degradation.
Without StyleGAN prior, the model predicts over-flattened results which lack of obvious 3D structure. Further, the predictions are over-smooth without displacement map. In contrast, the predictions of our method are fine-grained with details on eyebrow, eyelid and wrinkle.

**Textural Quality Evaluation:** We perform ablation study on textural modelling on our test set with high-resolution ground truth. The results are illustrated in Table 3. Our full method obtains best performance. In line (2), we observe that removing StyleGAN prior causes a significant decreasing of accuracy. Such results provide a consistent conclusion that L2R well propagates the facial clues. In lines (3)-(5), we observe the proposed modules in ARN also contribute to higher accuracy. Further, line (6) reveals better geometry modelling also improves the texture quality due to their mutual dependence in rendering, and line (7) reveals the effectiveness of our cross-view loss. We also perform visual comparison in Fig. 5. The proxy suffers from blurring of input images. Without StyleGAN prior, the quality of modelled texture is also limited. The predictions contain noise without GRN, which reveals the geometry modelling in L2R well boosts the texture reconstruction via rendering procedure. In contrast, our full model produces clearer and detailed texture.

**5.2. Comparison with State-of-the-Art Methods**

In this section, we compare L2R with recent state-of-the-art methods. Without specially statement, the methods use degraded image of $32 \times 32$ as input. To further analyse the difference between L2R and 2D super-resolution method, we use DFDNet [37] to pre-process the input image and denote such operation as ‘+DFDNet’. For methods using original image as input, we denote them with ‘origin’.

**Analysis on Geometry:** We first evaluate the predicted facial geometry on BFM dataset in Table 4, where the methods with tag ‘origin’ use original image as input without degradation. We observe that Unsup3D, GAN2Shape and LAP all suffer from low image quality which brings increase of geometry error. In contrast, L2R obtains sig-
Figure 6. Qualitative comparison of facial normal between our method and Unsup3D [58], LAP [67], DF2Net [63] and DECA [19]. We use DFDNet [37] to enhance the input appearance for other methods, by which we fairly compare with 2D pre-processed approaches. Our method obtains finer details on eyebrow, mustache and wrinkles, and keep smoothing on clean skins.

Figure 7. Visual comparison on texture against Unsup3D [58], LAP [67], Rotate-and-Render [68] and D3DFR [14]. All the compared methods are boosted by DFDNet [37] to enhance the input appearance, by which we fairly compare against 2D pre-processed approaches. Significantly superior performance in the same setting, which demonstrates the robustness of our method. Further, compared with line (3), (5) and (7), our method still outperforms the models without image quality limitation. We also make evaluation on 3DFAW dataset in Table 5, where line (0) represents the upper bound and line (1)-(4) provide ref-
erences from 3DMM-based or keypoint-estimation methods. We observe our L2R model obtains the best performance. Note that in line (6) and (9), although using DFDNet as 2D appearance pre-processing improves the accuracy to some extent, it still provides weaker performances than our method. Such phenomenon reveals that 2D appearance enhancement cannot well address image degradation for 3D modelling. We further analyse modelled facial normal on Photoface dataset in Table 6, where ‘-ft’ means finetuning on this dataset. Note that, ‘LAP origin’ uses original high-quality images as input, while Cross-Modal [3] approach inputs the original images and uses ground truth as supervision. In contrast, our method tackles a more challenging un-supervised setting of degraded input, and still obtains competitive results. Such analysis demonstrates the effectiveness of L2R method. Finally, we illustrate visual results in Fig. 6, where the other methods are enhanced by 2D super-resolution model DFDNet. For Unsup3D and LAP, we observe that 2D appearance enhancing cannot essentially improve their 3D modelling performance against degradation, yielding oversmooth geometry. For DF2Net and DECA, their results suffer from noise or incorrect facial structure. In contrast, our method successfully leverages the restored 2D facial clues for 3D modelling.

**Analysis on Texture:** We then perform evaluations on modelled texture on our high-resolution test set. For 3D texture completion method Rotate-and-Render, we only calculate cosine-similarity as its prediction is not spatially aligned with the ground truth. As illustrated in Table 7, 2D enhancement cannot significantly boost the modelled texture, while our L2R obtains best performance. Such phenomenon is consistent with previous experiments, which further demonstrates that our results have superior quality and cross-view consistency. Visual comparison is illustrated in Fig. 7. We find that although enhanced by DFDNet, Unsup3D and LAP cannot well complement the degradation of input image. The predictions of them lack of clear details and suffer from blurring. For Rotate-and-Render, its predictions also contain obvious noise and over-smoothness. For D3DFR, its performances are limited by the 3DMM basis and produces facial textures with lower reality. In contrast, our results have superior clarity, details, and less blurring.

### 5.3. Limitation

Here we analyse the potential limitation of our method. Several failure cases are shown in Fig. 8, in which we observe that our method may fail on faces with extreme poses, large expressions, artifacts or heavy make-ups. One possible reason is that the proxy also suffers from these challenging cases and produces corrupted geometry, thus our method is influenced by the proxy and cannot correctly align the final results. Another reason is about the data bias. Challenging cases including extreme poses, expressions or artifacts hardly appear in the CelebA or CelebAMask-HQ dataset. Hence, the model lacks experiences on addressing these factors. The third reason is due to the shape assumption. Without 3DMM, L2R has no reliable prior to correctly deal with these cases.

### 6. Conclusion and Discussion

In this paper, we propose a novel Learning to Restore 3D face (L2R) framework for high-quality 3D face modelling against image degradation. The core idea of L2R is transforming 2D generative facial priors to inherent 3D clues, and mutually boost 3D texture/geometry modelling over a simple proxy. To recover high-quality 3D texture, L2R constrains pre-trained StyleGAN with a novel albedo restoring network, which urges StyleGAN to provide facial priors of the required 3D position. To reconstruct detailed geometry, L2R leverages the restored 3D texture to improve explicit details modelling. Such two procedures are optimized progressively with a novel 3D-aware adversarial loss, yielding stable optimization and consistent prediction. In the future, several interesting directions could be considered beyond L2R method, e.g., leveraging explicit operations on StyleGAN rather than implicit transformation, or improving the approximated rendering operation. According to the discussion of limitation, non-parametric face modeling under challenging cases is also a possible direction.

**Broader Impact:** L2R predicts 3D faces based on the learned statistics of the training dataset. The potential biases may bring negative societal impacts. Besides, the model may generate inexisten contents. These issues warrant further research when building upon this work.

<table>
<thead>
<tr>
<th>No.</th>
<th>Method</th>
<th>Cosine-Similarity ↑</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Ours-full</td>
<td>0.722</td>
<td>0.885</td>
</tr>
<tr>
<td>(2)</td>
<td>Unsup3D [58]</td>
<td>0.582</td>
<td>0.495</td>
</tr>
<tr>
<td>(3)</td>
<td>Unsup3D + DFDNet</td>
<td>0.605</td>
<td>0.511</td>
</tr>
<tr>
<td>(4)</td>
<td>Unsup3D origin</td>
<td>0.627</td>
<td>0.520</td>
</tr>
<tr>
<td>(5)</td>
<td>LAP [7]</td>
<td>0.653</td>
<td>0.586</td>
</tr>
<tr>
<td>(6)</td>
<td>LAP + DFDNet</td>
<td>0.671</td>
<td>0.602</td>
</tr>
<tr>
<td>(7)</td>
<td>LAP origin</td>
<td>0.698</td>
<td>0.631</td>
</tr>
<tr>
<td>(8)</td>
<td>Rotate-and-Render [68]</td>
<td>0.611</td>
<td>-</td>
</tr>
<tr>
<td>(9)</td>
<td>Rotate-and-Render + DFDNet</td>
<td>0.640</td>
<td>-</td>
</tr>
<tr>
<td>(10)</td>
<td>Rotate-and-Render origin</td>
<td>0.697</td>
<td>-</td>
</tr>
<tr>
<td>(11)</td>
<td>D3DFR origin</td>
<td>0.798</td>
<td>0.335</td>
</tr>
</tbody>
</table>

Table 7. Texture evaluation with the state-of-the-arts.
References


Elad Richardson, Matan Sela, and Ron Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, pages 460–469, 2016.


