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

Zhenyu Zhang[1], Yanhao Ge[1], Ying Tai[1], Xiaoming Huang[1], Chengjie Wang[1]*,
Hao Tang[2], Dongjin Huang[3], Zhifeng Xie[3]*
Tencent Youtu Lab, Shanghai, China[1]
CVL, ETH Zurich, Switzerland[2]
Shanghai Film Academy of Shanghai University[3]
zhangjesse@foxmail.com hao.tang@vision.ee.ethz.ch
halege, yingtai, skyhuang, jasoncjwang@tencent.com djhuang, zhifeng_xie@shu.edu.cn

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) [8], which provides reasonable geometry assumptions for modelling. Based on this, the reconstruction can be achieved through optimization on low-dimensional parameters [46, 47]. With the development of deep learning, recent works utilize neural networks to regress 3DMM parameters from 2D images [45, 70]. Although 3DMM based approaches are further improved with non-linearity [19, 25, 53–55, 60, 69] and multi-view consistency [6, 10, 52, 57, 61], they still suffer from several drawbacks: limited amount of subjects (e.g., BFM [44] with 200 subjects) with controlled conditions, difficulties on building skin details, anatomic grounded muscles [16] and large variations of identity [71].

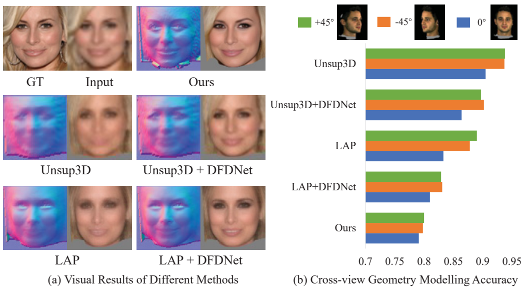

Figure 1. Experimental analyses among non-parametric methods. (a) Visual comparison under degraded input. (b) Cross-view scale-invariant depth error on MICC [5] dataset, where we use the models to predict geometry from the image of the current pose, and perform testing on images of other two poses. We use DFDNet [37] to preprocess the degraded images. Unsup3D [58] and LAP [67] suffer from the input degradation, and receive limited benefit from the 2D appearance enhancing method. In contrast, our method models more detailed predictions, and shows more robust performance in cross-view validation. Please refer to supplemental materials for more details.

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 [3, 49]. More recent works [42, 58] disentangle a face into intrinsic factors and accomplishes canonical reconstruction in an unsupervised manner via render-

*Chengjie Wang and Zhifeng Xie are corresponding authors 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 pretrained 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.

	Non-parametric	Supervision	3D Texture	Degraded Input
MOFA [54], DECA [19]	×	**I**, keypoint	✓	×
RingNet [48], MVF [57]	×	**I**, keypoint	×	×
D3DFR [14], GANfit [25]	×	**I**, keypoint	✓	✓
Cross-modal [3], DF2Net [63]	✓	**I**, 3D scan	×	×
Unsup3D [58], LAP [67]	✓	**I**	✓	×
Ours	✓	**I**	✓	✓

Table 1. Comparison with selected recent methods on settings. **I** means 2D image. Most methods do not tackle the condition of degraded input.

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 texture and geometry generation.

Unsup3D disentangles a facial image $\mathbf{I}$ into intrinsic factors (d, a, ω, l) including a canonical depth map $d \in \mathbb{R}_+$, 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 Λ and rasterization Π as follows:

$$\hat{\mathbf{I}} = \Pi(\Lambda(a, d, l), d, \omega). \tag{1}$$

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

$$\hat{\mathbf{I}}' = \Pi(\Lambda(a', d', l), d', \omega), \tag{2}$$

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

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}, \sigma) = -\frac{1}{|\Omega|}\sum \ln \frac{1}{\sqrt{2}\sigma} \exp - \frac{\sqrt{2}|\hat{\mathbf{I}} - \mathbf{I}|}{\sigma}, \tag{3}$$

where Ω is normalization factor. The flipped version $\mathcal{L}(\hat{\mathbf{I}}', \mathbf{I}, \sigma')$ 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, ω_o as the light/viewpoint proxy. The degraded input and high-resolution ground truth is denoted as $\mathbf{I}, \mathbf{I}_{gt}$.

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 $\mathbf{I}$, corresponding style code should be extracted to guide the pre-trained StyleGAN. We use a style encoder with the same architecture as those of Φ^d, Φ^a to get high-level features from $\mathbf{I}$, 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 $\mathbf{I}$, 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 Φ^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 the 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

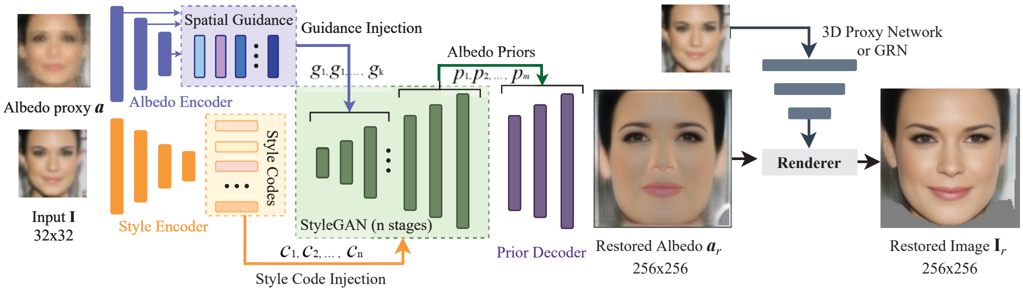

Figure 2. Overview of the proposed Albedo Restoration Network. Spatial guidance and style codes are extracted from original canonical albedo a_o and low-resolution image $\mathbf{I}$, respectively, and then fed to a pre-trained StyleGAN. Style code helps StyleGAN to generate priors with similar style 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 $\mathbf{I}_r$ can be rendered with the 3D proxy d_o, l_o and ω_o by Eqn. (1). The layers of StyleGAN are freezed during training.

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:

$$h_i = \begin{cases} Conv(p_i), & i = 1 \\ Conv([h_{i-1}, p_i]), & \text{otherwise} \end{cases} \tag{4}$$

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 $\mathbf{I_r}$ with the proxy of ω_o and depth d_o by Eqns. (1), (2). In our method, we recover albedo to a size of 256×256. The reason is that modelling 3D texture of higher resolution without supervision is challenging and causes 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 geometry from input $\mathbf{I}$, normal proxy n_o and restored canonical texture t_r. Here, features from t_r provide high-quality 3D texture clues in canonical view, while $\mathbf{I}$ 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

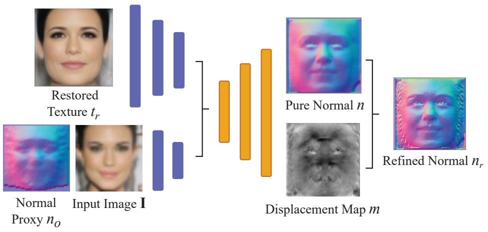

Figure 3. Overview of the Geometry Refining Network (GRN). An encoder-decoder network is utilized to predict pure normal n and displacement map m from input $\mathbf{I}$, normal proxy n_o and recovered canonical texture t_r. The refined geometry n_r is generated by Eqn. (5) for detailed 3D displacement. For better visualization, we show normal rather than depth.

we make it in a non-parametric manner. The displacement map $m \in [h, w, 3]$ is generated in the canonical view. Then, we project the learned pure depth d to 3D UV-space to get $D^{uv} \in [h, w, 3]$, and utilize m to enhance it as follows:

$$D_r^{uv} = D^{uv} + m \odot n, \tag{5}$$

where n is the normal map got from d. In this way, we get 3D detailed shape D_r^{uv} which can be utilized to enhance lighting Λ and reprojection Π in Eqn. (2).

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 $\dot{\omega}, \dot{l}$, and denote it as $\dot{\mathbf{I}}_r$. Given the high-resolution ground truth $\mathbf{I}_{gt}$, we use adversarial loss [4] with a same discriminator $\mathcal{D}$ as [33] to compare $\dot{\mathbf{I}}_r$ and $\mathbf{I}_{gt}$. This objective can be formulated as:

$$\mathcal{L}_c = \min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}[\log(\mathcal{D}(\mathbf{I}_{gt})] + \mathbb{E}[\log(1 - \mathcal{D}(\mathcal{G}(\dot{\mathbf{I}}_r))], \tag{6}$$

where $\mathcal{G}$ is our framework as generator. To suppress artifact in predicted displacement map, we use a Laplacian smoothness regularization which can be denoted as $\nabla|m|$. The final loss is then formulated as:

$$\mathcal{L}_{all} = \mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}_{gt}, \sigma) + \mathcal{L}(\hat{\mathbf{I}}', \mathbf{I}_{gt}, \sigma') + \alpha \mathcal{L}_c + \beta \nabla |m|, \quad (7)$$

where α, β 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 ω_o, l_o proxy predicted from Φ^ω, Φ^l. Then we jointly optimize ARN, GRN and Φ^ω, Φ^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 of 1024×1024, 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 $\mathbf{I_{gt}}$, and are further resized to 32×32 with different blurring and smoothing as degraded input $\mathbf{I}$. 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 Φ^ω, Φ^l as [58] to predict pose and light. For learning confidence map at output size of 256×256, we use a similar network Φ^σ 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. Further, 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.

No.	Method	SIDE ($\times 10^{-2}$) ↓	MAD (deg.) ↓
(1)	Ours-full	**0.710**±0.139	**14.70**±1.16
(2)	w/o StyleGAN prior	0.802±0.168	17.03±1.65
(3)	w/o StyleGAN prior + DFDNet [37]	0.729±0.145	15.14±1.52
(4)	w/o multi-style code	0.735±0.124	15.86±1.50
(5)	w/o spatial guidance	0.732±0.138	16.19±1.63
(6)	w/o prior decoder	0.728±0.129	15.63±1.58
(7)	w/o restored image	0.776±0.178	16.57±1.49
(8)	w/o displacement map	0.725±0.178	15.21±1.34
(9)	w/o cross-view loss	0.727±0.140	15.01±1.32

Table 2. Comparison with baselines on BFM dataset.

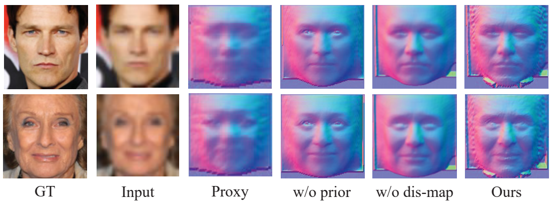

Figure 4. Ablation study on modeled geometry.

No.	Method	Cosine-Similarity ↑	SSIM ↑
(1)	Ours-full	**0.725**	**0.685**
(2)	w/o StyleGAN prior	0.631	0.528
(3)	w/o multi-style code	0.667	0.641
(4)	w/o spatial guidance	0.679	0.637
(5)	w/o prior decoder	0.690	0.644
(6)	w/o GRN	0.703	0.657
(7)	w/o cross-view loss	0.707	0.671

Table 3. Comparison with baselines on texture modelling.

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.

Texture: We then perform ablation study on texture 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 32×32 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'.

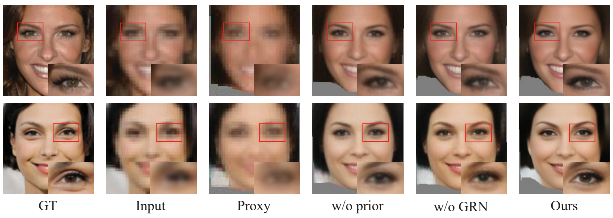

Figure 5. Ablation study on modelled texture.

No.	Method	SIDE ($\times 10^{-2}$) ↓	MAD (deg.) ↓
(1)	Ours	**0.710**±0.139	**14.70**±1.16
(2)	Unsup3D [58]	0.901±0.170	18.52±1.58
(3)	Unsup3D origin	0.793±0.140	16.51±1.56
(4)	GAN2Shape [42]	0.827±0.170	15.93±1.50
(5)	GAN2Shape origin	0.756±0.152	16.82±1.47
(6)	LAP [67]	0.856±0.142	16.77±1.33
(7)	LAP origin	0.721±0.128	15.53±1.42

Table 4. Results of state-of-the-art methods on BFM dataset.

No.	Method	Depth Corr. ↑	Time (ms)
(0)	Ground Truth	66	-
(1)	MOFA [54] (3DMM based)	15.97	-
(2)	D3DFR [14] (3DMM based)	50.05	-
(3)	DECA [19] (3DMM based)	51.93	-
(4)	DepthNet [41]	35.77	-
(5)	Unsup3D [58]	49.28	0.6
(6)	Unsup3D + DFDNet	52.43	0.6
(7)	Unsup3D origin	54.64	0.6
(8)	LAP [67]	51.48	2.0
(9)	LAP + DFDNet	56.25	2.0
(10)	LAP origin	57.92	2.0
(11)	Ours	**57.96**	1.6

Table 5. 3DFAW keypoint depth evaluation.

Method	MAD ↓	< 20° ↑	< 25° ↑	< 30° ↑
SfSNet [49]	25.5±9.3	43.6%	57.5%	68.7%
DF2Net [63] (GT)	24.3±5.7	42.2%	62.7%	74.5%
D3DFR [14]	23.5±6.1	46.1%	61.8%	73.3%
DECA [19]	**22.5**±5.3	48.7%	62.3%	73.7%
Cross-Modal [3] (GT)	22.8±6.5	**49.0**%	62.9%	74.1%
LAP [67] + DFDNet	24.2±5.6	47.3%	62.7%	74.5%
LAP origin	23.0±5.1	48.2%	63.1%	74.5%
Ours	23.2±**4.8**	48.4%	**63.5**%	**74.9**%
Cross-Modal-ft [3] (GT)	**12.0**±5.3	**85.2**%	92.0%	95.6%
LAP-ft origin [67]	12.3±4.5	84.9%	92.4%	96.3%
Ours-ft	12.5±**4.1**	85.0%	**92.5**%	**96.5**%

Table 6. Facial normal evaluation on Photoface dataset.

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-

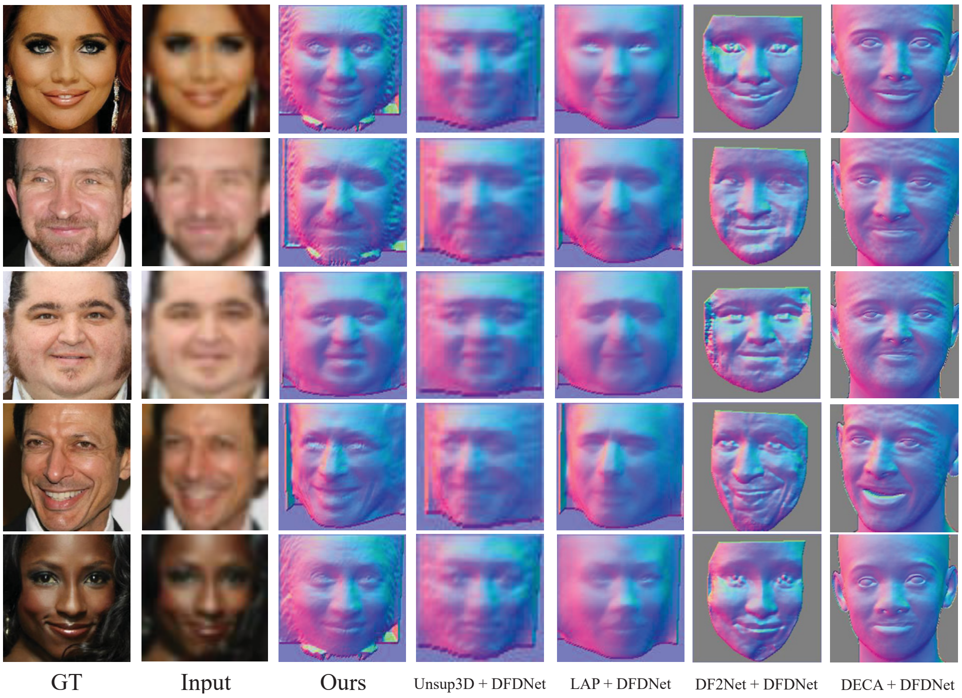

Figure 6. Qualitative comparison of facial normal between our method and Unsup3D [58], LAP [67], DF2Net [63] and DECA [19]. We use DFDNet [37] to enhance the input appearance for other methods, by which we fairly compare with 2D pre-processed approaches. Our method obtains finer details on eyebrow, mustache and wrinkles, and keep smoothing on clean skins.

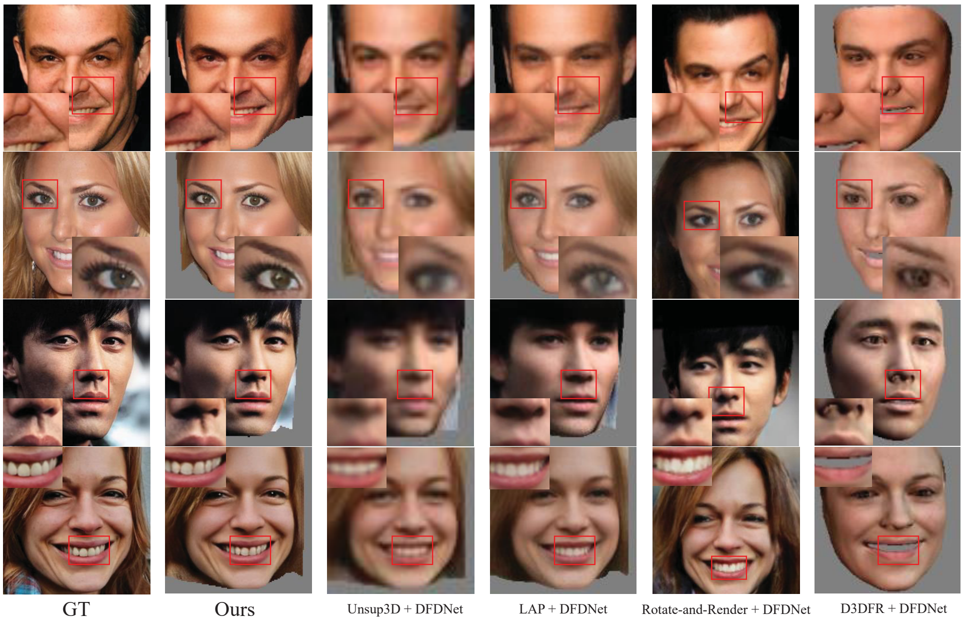

Figure 7. Visual comparison on texture against Unsup3D [58], LAP [67], Rotate-and-Render [68] and D3DFR [14]. All the compared methods are boosted by DFDNet [37] to enhance the input appearance, by which we fairly compare against 2D pre-processed approaches.

nificantly 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-

No.	Method	Cosine-Similarity ↑	SSIM ↑
(1)	Ours-full	**0.725**	**0.685**
(2)	Unsup3D [58]	0.582	0.493
(3)	Unsup3D + DFDNet	0.605	0.511
(4)	Unsup3D origin	0.627	0.520
(5)	LAP [67]	0.653	0.586
(6)	LAP + DFDNet	0.671	0.602
(7)	LAP origin	0.698	0.631
(8)	Rotate-and-Render [68]	0.611	-
(9)	Rotate-and-Render + DFDNet	0.640	-
(10)	Rotate-and-Render origin	0.697	-
(11)	D3DFR origin	0.398	0.335

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

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 unsupervised 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 predicts accurate geometry with details on eyebrow and wrinkles, which reveals L2R successfully leverages 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.

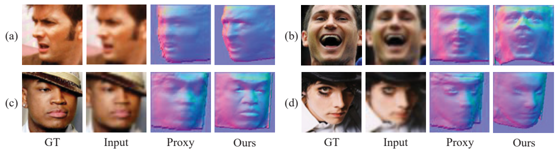

Figure 8. Failure cases of L2R. (a) Extreme poses. (b) Extreme expressions. (c) Large artifacts. (d) Heavy make-ups or shadows.

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 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 inexistent contents. These issues warrant further research when building upon this work.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *CVPR*, pages 4432–4441, 2019. 3

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, pages 8296–8305, 2020. 3

[3] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *CVPR*, pages 4979–4989, 2020. 1, 2, 5, 6, 8

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv: 1701.07875*, 2017. 4

[5] 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. 1

[6] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *CVPR*, pages 5850–5860, 2020. 1

[7] David Bau, Hendrik Strobelt, William Peebles, Bolei Zhou, Jun-Yan Zhu, Antonio Torralba, et al. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. 3

[8] 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, 2

[9] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. *arXiv preprint arXiv:2012.00739*, 2020. 3

[10] Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized face modeling for improved face reconstruction and motion retargeting. In *ECCV*, 2020. 1

[11] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *ICCV*, pages 9429–9439, 2019. 4

[12] 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*, pages 7093–7102, 2018. 2

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 5

[14] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 2019. 2, 6, 7

[15] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5908–5917, 2017. 2

[16] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 1

[17] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, pages 2366–2374, 2014. 5

[18] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *arXiv preprint arXiv:2012.04012*, 2020. 4

[19] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1, 2, 6, 7

[20] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, pages 534–551, 2018. 2

[21] Stephan J Garbin, Marek Kowalski, Matthew Johnson, and Jamie Shotton. High resolution zero-shot domain adaptation of synthetically rendered face images. In *ECCV*, pages 220–236, 2020. 2

[22] Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised adversarial learning to generate photo realistic face images of new identities from 3d morphable model. In *ECCV*, pages 217–234, 2018. 2

[23] Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *CVPR*, pages 7628–7638, 2021. 2

[24] Baris Gecer, Alexandros Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In *ECCV*, pages 415–433, 2020. 2

[25] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, pages 1155–1164, 2019. 1, 2

[26] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, pages 8377–8386, 2018. 2

[27] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 5

[28] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, pages 3012–3021, 2020. 3

[29] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 3

[30] 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*, pages 1031–1039, 2017. 2

[31] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *IEEE international conference and workshops on automatic face and gesture recognition*, volume 1, pages 1–8, 2015. 5

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 3, 5

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2, 3, 4, 5

[34] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018. 2, 3

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[36] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 5

[37] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, pages 399–415, 2020. 1, 2, 5, 6, 7

[38] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *ECCV*, pages 545–560. Springer, 2016. 2

[39] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 5

[40] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, pages 2437–2445, 2020. 2

[41] Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, and Chris Pal. Unsupervised depth estimation, 3d face rotation and replacement. In *NeurIPS*, pages 9736–9746, 2018. 6

[42] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *arXiv preprint arXiv:2011.00844*, 2020. 1, 2, 6

[43] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, pages 262–277, 2020. 3

[44] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 1, 5

[45] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *3DV*, pages 460–469, 2016. 1, 2

[46] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *Computer Vision*, page 59. IEEE, 2003. 1, 2

[47] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, volume 2, pages 986–993, 2005. 1

[48] 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 *CVPR*, pages 7763–7772, 2019. 2

[49] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In *CVPR*, pages 6296–6305, 2018. 1, 2, 5, 6

[50] Gil Shamai, Ron Slossberg, and Ron Kimmel. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(3s):1–24, 2019. 2

[51] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *CVPR*, pages 6258–6266, 2021. 2

[52] 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 *CVPR*, pages 10812–10822, 2019. 1, 2

[53] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *CVPR*, pages 3361–3371, 2021. 1, 2

[54] 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 *ICCVW*, pages 1274–1283, 2017. 1, 2, 6

[55] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, pages 7346–7355, 2018. 1

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 5

[57] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, pages 959–968, 2019. 1, 2

[58] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, pages 1–10, 2020. 1, 2, 3, 5, 6, 7, 8

[59] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *CVPR*, pages 601–610, 2020. 4

[60] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *CVPR*, pages 12803–12813, 2021. 1, 2

[61] Jae Shin Yoon, Takaaki Shiratori, Shoou-I Yu, and Hyun Soo Park. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In *CVPR*, pages 4601–4609, 2019. 1

[62] Stefanos Zafeiriou, Mark Hansen, Gary Atkinson, Vasileios Argyriou, Maria Petrou, Melvyn Smith, and Lyndon Smith.

The photoface database. In *CVPRW*, pages 132–139, 2011. 5

[63] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *ICCV*, pages 2315–2324, 2019. 2, 6, 7

[64] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(8):690–706, 1999. 2

[65] 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 *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013. 5

[66] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 5

[67] Zhenyu Zhang, Yanhao Ge, Renwang Chen, Ying Tai, Yan Yan, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning to aggregate and personalize 3d face from in-the-wild photo collection. In *CVPR*, pages 14214–14224, 2021. 1, 2, 5, 6, 7, 8

[68] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photo realistic face rotation from single-view images. In *CVPR*, pages 5911–5920, 2020. 2, 7, 8

[69] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *CVPR*, pages 1097–1106, 2019. 1, 2

[70] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016. 1, 2

[71] Xiangyu Zhu, Fan Yang, Chang Yu Di Huang, Hao Wang, Jianzhu Guo, Zhen Lei, and Stan Z Li. Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In *ECCV*, 2020. 1