UV-GAN: Adversarial Facial UV Map Completion for Pose-invariant Face Recognition

Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, Stefanos Zafeiriou
Imperial College London
{j.deng16, shiyang.cheng11,n.xue15,yuxiang.zhou10,s.zafeiriou}@imperial.ac.uk

Abstract

Recently proposed robust 3D face alignment methods establish either dense or sparse correspondence between a 3D face model and a 2D facial image. The use of these methods presents new challenges as well as opportunities for facial texture analysis. In particular, by sampling the image using the fitted model, a facial UV can be created. Unfortunately, due to self-occlusion, such a UV map is always incomplete. In this paper, we propose a framework for training Deep Convolutional Neural Network (DCNN) to complete the facial UV map extracted from in-the-wild images. To this end, we first gather complete UV maps by fitting a 3D Morphable Model (3DMM) to various multiview image and video datasets, as well as leveraging on a new 3D dataset with over 3,000 identities. Second, we devise a meticulously designed architecture that combines local and global adversarial DCNNs to learn an identity-preserving facial UV completion model. We demonstrate that by attaching the completed UV to the fitted mesh and generating instances of arbitrary poses, we can increase pose variations for training deep face recognition/verification models, and minimise pose discrepancy during testing, which lead to better performance. Experiments on both controlled and in-the-wild UV datasets prove the effectiveness of our adversarial UV completion model. We achieve state-of-the-art verification accuracy, 94.05%*, under the CFP frontal-profile protocol only by combining pose augmentation during training and pose discrepancy reduction during testing. We will release the first in-the-wild UV dataset (we refer as WildUV) that comprises of complete facial UV maps from 1,892 identities for research purposes.*

1. Introduction

During the past few years, we have witnessed considerable progress in sparse and dense 3D face alignment. Some of the developments include the use of Deep Neural Networks (DNNs) to recover 3D facial structure [30, 47, 21], as well as a robust framework for fitting a 3D Morphable Model (3DMM) to images in-the-wild [3]. Furthermore, benchmarks suitable for training sparse 3D face alignment models have been developed recently [44, 20]. The utilisation of these methods introduces new challenges and opportunities, as far as facial texture is concerned. More specifically, by sampling over the fitted image, a facial UV map that embeds the manifold of a 3D face as a 2D contiguous atlas can be created. An example of the facial UV map is shown in Fig. 1. It is evident that such UV map contains a considerable amount of missing pixels due to self-occlusion (filled with random noise in the figure). In this paper, we tackle the problem of facial UV map completion from a single image, which has not received considerable attention. We demonstrate the usefulness of the proposed UV completion framework by creating synthetic samples to train deep neural networks for face recognition.

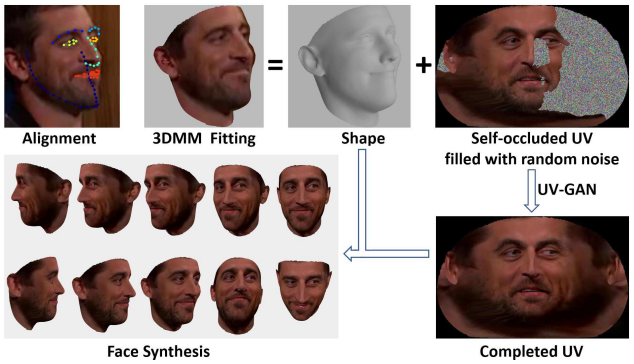


Figure 1. Adversarial UV completion. After fitting a 3DMM to the image, we retrieve a 3D face with an incomplete UV map. We learn a generative model to recover the self-occluded regions. By rotating a 3D shape with complete UV map, we can generate 2D faces of arbitrary poses, which can either augment pose variations during training or narrow pose discrepancy during testing for pose-invariant face recognition.

Face representation using Deep Convolutional Neural Network (DCNN) embeddings is considered the method of choice for face verification, face clustering, and recognition [35, 32, 26, 34]. However, when it comes to

frontal-profile face verification, performance of most DCNNs drops drastically by over 10% [33] compared with frontal-frontal verification, while human performance only degrades slightly. This indicates that the pose variation remains a significant challenge in face recognition. One approach to learn pose-invariant discriminative representations is to collect training data with large pose variations. However, web face images usually have long-tailed pose distributions [24] and it is not feasible to collect and label data that provide good coverage for full poses for all identities.

In this paper, we propose an adversarial UV completion framework (UV-GAN) and apply it to solve the pose-invariant face recognition problem without the requirement of extensive pose coverage in training data. Our framework is depicted in Fig. 1, we first fit a 3DMM to 2D image and retrieve the incomplete facial UV. To complete the UV map, we combine local and global adversarial networks to learn identity-preserving full UV texture (Sec. 3.2). We attach the complete UV with the fitted mesh and generate synthetic data with arbitrary poses. To enlarge the diversity of training images under various poses (along with pose labels), at no additional labelling cost, we synthesise a large number of profile images from the CASIA dataset [42]. We also use such synthesis method to reduce the pose discrepancy during testing. Our frontal-profile verification results show that rendering frontal face to profile view obtains better performance than frontalising profile face. In contrast to existing face frontalisation methods [43, 17], the proposed method is more effective and flexible for pose-invariant face recognition.

To summarise, our key contributions are:

- We are the first to apply local, global and identity-preserving adversarial networks to the problem of UV map completion. We show that the proposed method can generate realistic and coherent UV maps under both controlled and in-the-wild settings even when the missing regions account for 50% of the UV map. Using the completed UV and the corresponding 3D shape, we are able to synthesise 2D face images with arbitrary poses.
- For face recognition training, our pose synthesis method can enrich the pose variations of training data, without incurring the expense of manually labelling large datasets spanning all poses. For recognition testing, our method can narrow the pose discrepancy between the verification pairs resulting in better performance. We obtain state-of-the-art verification accuracy, 94.05%, under the CFP frontal-profile protocol.
- To the best of our knowledge, we are the first to collect a large-scale dataset of complete ear-to-ear UV facial maps of both controlled and in-the-wild data. We will release the first in-the-wild UV dataset (referred as WildUV) that comprises of complete facial UV maps from 1,892 identities for research purposes.

2. Related Work

Image Completion Image completion has been studied in numerous contexts, including inpainting and texture synthesis. Pathak *et al*. [27] propose context encoders with a reconstruction and an adversarial loss to generate the contents for missing regions that comply with the neighbourhood regions. Yang *et al*. [41] further improve inpainting with a multi-scale neural patch synthesis method. This approach is based on a joint optimisation of image content and texture constraints, which not only preserves contextual structures but also produces fine details. Possible architectures, based on deep neural networks, for image completion are the so-called Pixel Recurrent Neural Networks (Pixel-RNNs) [25] and Pixel CNNs [38]. Nevertheless, we found that these architectures are mainly suitable for low-resolution images, whereas we want to design a DCNN architecture that can handle high resolutions facial UV maps. Probably the closest work to ours is the face completion method in [22]. This method combines a reconstruction loss, two adversarial losses and a semantic parsing loss to ensure genuineness and consistency of local-global contents [18]. However, their method is trained only on a small set of predefined masks and is not directly applicable to our problem, since (a) each mask has semantical correspondence with a 3D face, and (b) the missing regions of profile mask may take up over 50% of the image.

Pose-invariant Face Recognition Recent DCNNs [35, 32, 26, 34] trained on large-scale datasets have significantly boosted the performance of face recognition, the robustness of these methods against pose variations is sourced from the training data. To learn pose-invariant feature embedding, Masi *et al*. [24] and Zhao *et al*. [46] synthesise face image appearances across 3D viewpoints to increase pose variations in the training data. Peng [28] further explore reconstruction-based disentanglement during training for pose-invariant face recognition. With the introduction of GAN, pose-invariant feature disentanglement [36, 37] and face frontalisation [43, 17] methods become quite popular. During testing, frontal faces are generated from the generator, which decreases the pose discrepancy and henceforth improves pose-invariant face recognition. However, those approaches usually rely on large amount of pairings across poses, which is over demanding under in-the-wild scenario. By contrast, we propose an adversarial UV completion method that enables us to leverage frontal and four-quarter face pairs for synthesising training examples close to the ground truth.

3. Proposed Approach

In this section, we describe the proposed model for UV completion. In brief, after a 3DMM fitting on the 2D face image, we sample the image and retrieve a UV texture for the 3D mesh. Due to face self-occlusion, such UV texture is always incomplete. Hence, our goal is to synthesise the missing contents of the UV map that are semantically consistent with the whole UV map as well as visually realistic and identity-preserved. Fig. 2 illustrates the proposed network consisting of one generator and two discriminators.

3.1. 3D Morphable Model Fitting

To recover a 3D face from the 2D image, 3DMM fitting is employed. We adopt the 3DMM fitting proposed in [3] for our task. In [3], three parametric models are defined and need to be solved: *shape* (Eq. 1), *texture* (Eq. 2) and *camera* (Eq. 3) models:

$$\mathcal{S}(\mathbf{p}) = \bar{\mathbf{s}} + \mathbf{U}_s\mathbf{p}, \tag{1}$$

$$\mathcal{T}(\boldsymbol{\lambda}) = \bar{\mathbf{t}} + \mathbf{U}_t\boldsymbol{\lambda}, \tag{2}$$

$$\mathcal{W}(\mathbf{p}, \mathbf{c}) = \mathcal{P}(\mathcal{S}(\mathbf{p}), \mathbf{c}), \tag{3}$$

where $\mathbf{p}, \boldsymbol{\lambda}$ and $\mathbf{c}$ are shape, texture and camera parameters to optimise. $\mathbf{U}_s$ and $\mathbf{U}_t$ are the shape and texture eigenbasis respectively, $\bar{\mathbf{s}}$ and $\bar{\mathbf{t}}$ are mean shapes of shape and texture models correspondingly, which are learnt from $10,000$ face scans of different individuals [4]. Function $\mathcal{P}$ is a perspective camera transformation. Thus the overall cost function for 3DMM fitting is formulated as:

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda},\mathbf{c}} \|\mathbf{F}(\mathcal{W}(\mathbf{p},\mathbf{c})) - \mathcal{T}(\boldsymbol{\lambda})\|^2 + \alpha_s\|\mathbf{p}\|^2_{\boldsymbol{\Sigma}_s^{-1}} + \alpha_t\|\boldsymbol{\lambda}\|^2_{\boldsymbol{\Sigma}_t^{-1}}. \tag{4}$$

Here, $\|\mathbf{p}\|^2_{\boldsymbol{\Sigma}_s^{-1}}$ and $\|\boldsymbol{\lambda}\|^2_{\boldsymbol{\Sigma}_t^{-1}}$ are two regularisation terms to counter over-fitting, $\boldsymbol{\Sigma}_s^{-1}$ and $\boldsymbol{\Sigma}_t^{-1}$ are diagonal matrices with the main diagonal being eigenvalues for the shape and texture model respectively. α_s and α_t are constants empirically set to weigh the two regularisation terms. Note that $\mathbf{F}(\mathcal{W}(\mathbf{p},\mathbf{c}))$ denotes the operation of sampling the feature-based input image on the projected 2D locations.

To accelerate the 3DMM fitting of the face images in the wild, we initialise the fitting with 3D landmarks[1] predicted by [5], and utilise the landmarks subsequently in the optimisation by adding a 2D landmark term in Eq. 4. The final objective function could be written as:

$$\arg\min_{\mathbf{p},\boldsymbol{\lambda},\mathbf{c}} \|\mathbf{F}(\mathcal{W}(\mathbf{p},\mathbf{c})) - \mathcal{T}(\boldsymbol{\lambda})\|^2 + \alpha_l\|\mathcal{W}_l(\mathbf{p},\mathbf{c}) - \mathbf{s}_l\|^2 + \alpha_s\|\mathbf{p}\|^2_{\boldsymbol{\Sigma}_s^{-1}} + \alpha_t\|\boldsymbol{\lambda}\|^2_{\boldsymbol{\Sigma}_t^{-1}}, \tag{5}$$

where $\mathbf{s}_l$ is the 2D shape, α_l is a constant to weigh the landmark term. Eq. 5 could be solved by the Gauss-Newton optimisation framework in [3].

[1]The 3D landmarks here refer to the 2D projections of the 3D facial landmarks [44, 10, 9].

Based on fitting results, we categorise face images from CASIA dataset [42] into 13 pose groups [37]. In Fig. 3, the mean face from [37] becomes more blurred as the yaw angle increases. By contrast, our mean faces remain clear across all pose groups, which indicates that our fitting method produces much more accurate poses.

3.2. UV Texture Completion

Once we have the estimated 3D shape of the facial image, its visible vertices could be computed by z-buffering, and then utilised to generate a visibility mask for UV texture. For those *invisible* parts (*i.e.* missing textures), we want to fill them with the identity preserving textures. We propose a Generative Adversarial Network for UV completion (we refer as **UV-GAN**), it comprises of one UV generation module, two discriminators and an additional module to preserve face identity.

Generation Module Given input UV textures with missing regions, the generator G works as an auto-encoder to construct new instances. We adopt pixel-wise l_1 norm as the reconstruction loss:

$$L_{gen} = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\left|I_{i,j} - I^*_{i,j}\right|, \tag{6}$$

where $I_{i,j}$ is the estimated UV texture and $I^*_{i,j}$ is the ground truth texture. To preserve the image information in original resolution, we follow the encoder-decoder design in [19], where skip connections between mirrored layers in the encoder and decoder stacks are made. We fill the incomplete UV texture with random noise and concatenate with its mirror image as the generator input. Since the face is not exactly symmetric, we avoid using symmetry loss as in [17]. Unlike the original GAN model [12] that initialises from a noise vector, the hidden representations obtained from our encoder capture more variations as well as relationships between invisible and visible regions, and thus help the decoder to fill up missing regions.

Discrimination Module Despite that previous generation module could fill missing pixels with small reconstruction errors, it does not guarantee the output textures to be visually realistic and informative. In Fig. 5 (a), we show examples of the generated texture from the generator, which are quite blurry and missing important details. To improve the quality of synthetic images and encourage more photo-realistic results, we adopt a discrimination module D to distinguish real and fake UVs. The adversarial loss, which is a reflection of how the generator could maximally fool the discriminator and how well the discriminator could distinguish between real and fake, is defined as:

$$L_{adv} = \min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_d(\mathbf{x}),\mathbf{y}\sim p_d(\mathbf{y})}\left[\log D(\mathbf{x},\mathbf{y})\right] + \mathbb{E}_{\mathbf{x}\sim p_d(\mathbf{x}),\mathbf{z}\sim p_z(\mathbf{z})}\left[\log(1 - D(\mathbf{x}, G(\mathbf{x},\mathbf{z})))\right], \tag{7}$$

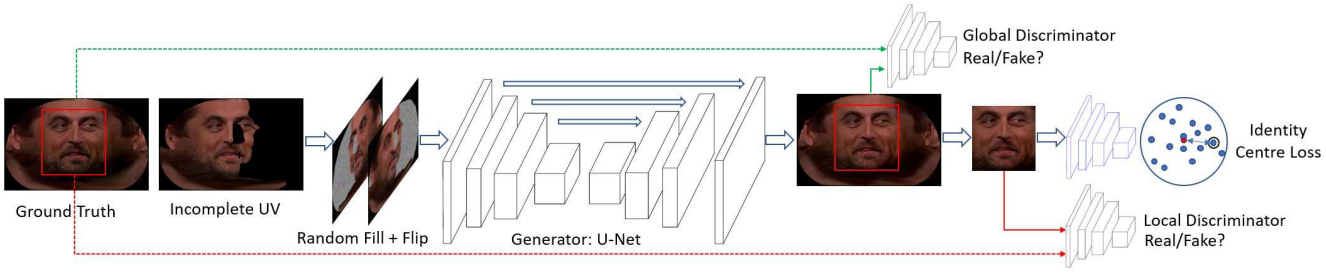


Figure 2. Network architecture. It consists of one generator and two discriminators. The generator takes the incomplete UV map and its flip filled with random noise as input and outputs the full UV map. Two discriminators are learned to validate the genuineness of the synthetic UV texture and the main face region. The pre-trained identity classification network that would always be fixed, is to further ensure the newly generated faces preserve the identity. Note that only the generator is required for the testing stage.

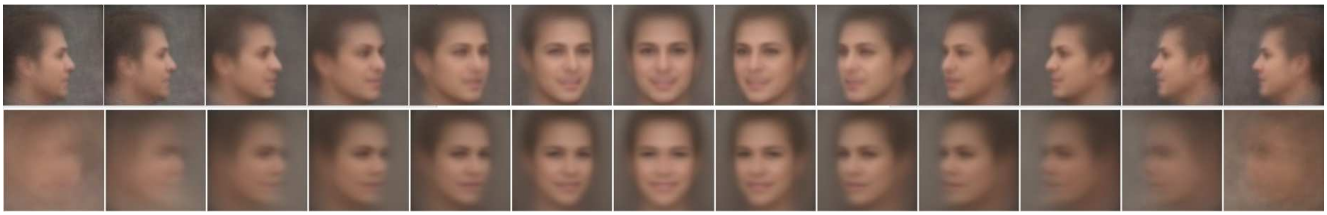


Figure 3. The mean faces of 13 pose groups in the CASIA dataset [42]. The first row is our results, and the second row is from [37]. Our mean faces are more clear under large pose variations, and there is only slight blurriness on facial organs even for profile faces.

where $p_z(\mathbf{z})$, $p_d(\mathbf{x})$ and $p_d(\mathbf{y})$ represent the distributions (Gaussian) of the noise variable $\mathbf{z}$, the partial UV texture $\mathbf{x}$ and the full UV texture $\mathbf{y}$ respectively.

Particularly, our module D consists of two discriminators: a global and a local discriminator. We firstly train a global discriminator to determine the faithfulness of the entire UV maps. The core idea is that synthetic contents should not only look realistic, but also conform to their surrounding contexts. Furthermore, we define a local discriminator that focuses on the face centre. There are two reasons for introducing the local discriminator: (1) for UV faces in the wild, outer face (*e.g.* ears, forehead) is usually noisy and unreliable; (2) inner face is considered much more informative as fas as identity is concerned. Compared with the global discriminator, the local module (Fig. 5 (d)) enhances the central face region with less noisy and sharper boundaries. The benefit of combining global and local discriminator is: the global one maintains the context of the facial image, while the local discriminator enforces the generated texture to be more informative within central face region. Similar GAN architecture with two discriminators could be found in [22, 17].

Identity Preserving Module Preserving the identity while synthesising faces is the most critical part in developing the recognition-via-generation framework. We exploit the centre loss [40] to improve the identity preserving ability of our UV-GAN. Specifically, we define the centre loss based on the activations after the average pooling layer of ResNet-27 [40, 15].

$$L_{id} = \frac{1}{m}\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{c}_{y_i}\|_2^2, \tag{8}$$

where m is the batch size, $\mathbf{x}_i \in \mathbb{R}^{512}$ is the embedding features and $\mathbf{c}_{y_i} \in \mathbb{R}^{512}$ denotes the y_i-th class feature centre. Note that ResNet-27 is pre-trained on CASIA [42] dataset using the softmax loss to classify 10k identities. It captures the most prominent feature and facial structure to discriminate identity. Hence, it is beneficial to leverage this loss to maintain identity in the synthetic texture. As our feature embedding network is fixed during training, all the generated samples would lie close to their own fixed feature centre.

Objective Function The final loss function for the proposed UV-GAN is a weighted sum of aforementioned losses:

$$L = L_{gen} + \lambda_1 L_{adv_g} + \lambda_2 L_{adv_l} + \lambda_3 L_{id}, \tag{9}$$

where L_{adv_g} and L_{adv_l} are the global and local adversarial losses as in Eq. 7, and λ_1, λ_2 and λ_3 are the weights to balance different losses.

4. Experiments

4.1. Settings and Datasets

For UV completion, the original size of ground truth UV maps is 377×595 and we re-scale the incomplete UV maps to 256×256 as the input of the UV-GAN model. The network structures of the generator and discriminator follow[2] [19]. The input of the local discriminator is a 128×128 crop on facial feature regions as shown in Fig 2, and local discriminator has the same structure as the global discriminator. For each complete facial UV map used for training,

[2]https://github.com/phillipi/pix2pix

we create 20 synthetic 2D facial images with arbitrary yaw angles $[-90°, 90°]$, compute the visibility masks and use the corresponding incomplete facial UV maps for training. Our network is implemented with Tensorflow [1]. We train UV-GAN for 100 epochs with a batch size of 8 and a learning rate of 10^{-4}. In all our experiments, we empirically set $\lambda_1 = 10^{-2}$, $\lambda_2 = 4 \times 10^{-2}$, and $\lambda_3 = 10^{-3}$.

UVDB. UVDB is a dataset that we developed for training the proposed UV-GAN. It has been built from three different sources. The first subset contains 3,564 subjects (3,564 unique identities with six expressions, 21,384 unique UV maps in total) scanned by the 3dMD device[3]. The second subset is constructed from Multi-PIE [13] with 337 identities (2514 unique facial UV maps for each illumination setting, 50,280 in total). The third subset (called **WildUV**) is constructed from UMD video dataset [2]. We have selected videos with large pose variations to get coverage in all different poses and finally developed the first in-the-wild UV dataset with 1,892 identities (5,638 unique UV maps). Poisson blending [29] was used for improving the quality of the facial UV maps. As is shown in Fig. 4, the UVDB contains facial UV maps captured in controlled conditions, as well as challenging in-the-wild videos (*e.g.*, low resolution with potential occlusions). We will release WildUV subset with 2D face images, 3D fitting results and the complete UV maps. For the 3dMD and WildUV subsets, the first 90% of identities are used for training and the rest are used for testing. For the MultiPIE subset, the first 200 subjects are used for training and the rest 137 subjects for testing [36, 37]. For face synthesis, we use all the training subsets of UVDB.

Figure 4. UV Completion Dataset.

CASIA [42] contains 494,414 images of 10,575 subjects.

VGG2 [6] consists of a training set with 8,631 identities (3,141,890 images) and a test set with 500 identities (169,396 images).

MS1M [14] includes about 100k identities with 10 million images. In this paper, we use a refined version [11, 8] of MS1M.

[3]http://www.3dmd.com/

CFP [33] consists of 500 subjects, each with 10 frontal and 4 profile images. The evaluation protocol includes frontal-frontal (FF) and frontal-profile (FP) face verification, each having 10 folders with 350 same-person pairs and 350 different-person pairs.

4.2. UV Completion

To quantitatively evaluate the UV completion results, we employ two metrics. The first one is the peak signal-to-noise ratio (PSNR) which directly measures the difference in pixel values. The second one is the structural similarity index (SSIM) [39] that estimates the holistic similarity between two images. These two metrics are computed between the predicted UV maps and the ground truth.

UVDB	(a)	(b)	(c)	(d)	(e)	CE
3dMD	25.8 0.889	26.3 0.895	25.2 0.879	25.7 0.886	26.5 0.898	25.1 0.856
MultiPIE	25.2 0.881	25.7 0.885	24.6 0.865	25.2 0.873	25.8 0.886	24.5 0.842
WildUV	22.3 0.872	22.8 0.876	22.0 0.861	22.5 0.868	22.9 0.887	21.6 0.840

Table 1. Quantitative evaluations of profile UV completion under different settings of our model and Context Encoder [27]. (a): L_{gen}. (b): flip + L_{gen}. (c): flip + L_{gen} + L_{adv_g}. (d) flip + L_{gen} + L_{adv_g} + L_{adv_l}. (e) UV-GAN. For each subset, the first row is PSNR value (dB) and the second row is SSIM value.

We first conduct the ablation study of the proposed UV-GAN under different settings and also compare with Context Encoder [27]. Fig. 5 shows UV completion results for profile faces on the UVDB. We have also zoomed in on the occluded eye to more clearly reveal the sharpness and authenticity of the recovery result by our combined loss. With only reconstruction loss L_{gen}, the completed UV maps are smooth and blurry. When adding flipped faces in training, the model converges faster and leads to better results (Tab. 1), but the completed images are still blurry. With global and local adversarial losses (L_{adv_g} + L_{adv_l}), the UV completion results look much more visually realistic and coherent globally and locally. Note that incorporating global adversarial loss slightly decreases PSNR and SSIM values (see Tab. 1 (c)), which are observed similarly in generative face completion [22]. Notwithstanding, when coupled with local and identity-preserving modules, such effects have been mitigated (see Tab. 1 (d) & (e)). The closest to our method is the CE method proposed in [27]. For a fair comparison, we have trained the CE method with our data. As can be seen, the proposed UV-GAN performs consistently better than the CE model both qualitatively and quantitatively.

Fig. 6 shows the performance of the proposed UV-GAN model under different poses from $0°$ to $90°$ in intervals of $15°$ on the UVDB. The completion results on the controlled

Figure 5. UV Completion results for profile faces on UVDB under different settings of our model and Context Encoder [27]. (a): L_{gen}. (b): flip + L_{gen}. (c): flip + $L_{gen} + L_{adv_g}$. (d) flip + $L_{gen} + L_{adv_g} + L_{adv_l}$. (e) UV-GAN. The proposed UV-GAN shows the most realistic and plausible completed content.

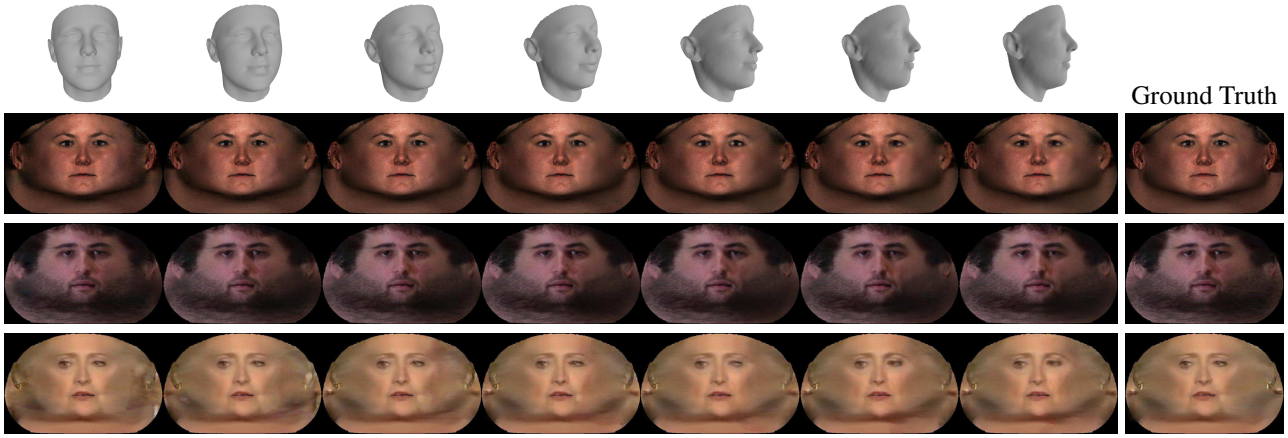


Figure 6. UV completion results under view changes on UVDB. Please zoom in to check the completion results on the self-occluded facial parts (*e.g.* eye regions).

UVDB	Metric	0°	±30°	±60°	±90°
3dMD	PSNR	28.3	27.9	27.0	26.5
	SSIM	0.920	0.907	0.903	0.898
MPIE	PSNR	27.5	26.7	26.5	25.8
	SSIM	0.912	0.903	0.899	0.886
WildUV	PSNR	26.3	25.2	24.5	22.9
	SSIM	0.908	0.902	0.894	0.887

Table 2. Quantitative evaluations of UV-GAN under view changes.

and in-the-wild faces are both visually realistic and consistent. In the profile case, we find that the only imperfection is the slight blurriness on the self-occluded eyes, which indicates that the proposed UV-GAN can successfully complete the UV maps in side views. In Tab. 2, we observe that the completion quality deteriorates as the yaw angle increases because the self-occluded regions are enlarging. Moreover, the completion results degrade as the data change from highly controlled 3dMD data to totally in-the-wild data, which indicates that UV completion in-the-wild is more challenging perhaps due to low resolution and occlusions.

4.3. Pose-invariant Face Recognition

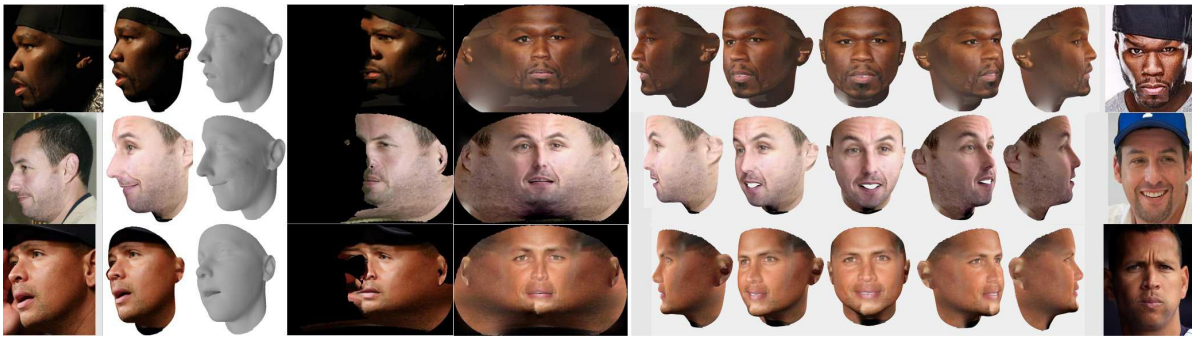
We pre-process all the face images by applying the face detection [45] and a 3DMM fitting. After that, face images are cropped to 112×112. To study the impacts of training data and loss function on recognition, we use the same network architecture in all our experiments (ResNet-27[4] [40]) without any bells and whistles. That is, we only vary the training data and the loss functions, which include centre loss [40], sphere loss [23] and softmax loss.

We abbreviate each training strategy as [dataset]-[loss]-[aug]. For example, CASIA-sm means that we have trained the network using CASIA dataset with the soft-max loss (denoted by sm), while CASIA-sphere means that we have used the sphere loss. CASIA-sm-aug1 refers to training the network in the CASIA dataset using the softmax loss (sm) and the augmentation method proposed in [24]. CASIA-sm-augUV refers to training using the same dataset and loss as above but augmenting the data using UV-GAN. Our augmentation method is as follows: we create 300 face images per identity (similar to Fig. 7). For frontal faces in the train-

[4]https://github.com/ydwen/caffe-face

Method	LFW	0°- 0°	0°- 45°	0°- 90°	45°- 0°	45°- 45°	45°- 90°	90°- 0°	90°- 45°	90°- 90°
CASIA-center	99.10	0.8128	0.8001	0.7018	0.8088	0.8199	0.7328	0.7081	0.7332	0.7331
CASIA-sphere	99.27	0.8010	0.7678	0.6211	0.7784	0.7934	0.6708	0.6262	0.6749	0.6957
CASIA-sm	99.02	0.8158	0.7878	0.6483	0.7946	0.8135	0.6955	0.6515	0.7015	0.7233
CASIA-sm-augUV	99.22	0.8342	0.8182	0.7237	0.8256	0.8404	0.7582	0.7302	0.7597	0.7682
VGG2-sm	99.35	0.8397	0.8262	**0.7325**	0.8319	0.8486	**0.7672**	**0.7386**	**0.7704**	**0.7805**
MS1M-sm	**99.60**	**0.8605**	**0.8427**	0.6693	**0.8486**	**0.8572**	0.6993	0.6718	0.7050	0.7172

Table 3. Face probing across poses on VGG2 test set. Accuracy on LFW [16] is also put in the left as a performance reference. Cosine similarity scores are evaluated across pose templates. A higher value is better.



(a) Profile Face UV Completion

(b) Frontal Face UV Completion

Figure 7. Profile face and frontal face UV Completion on CFP dataset. From left to right: 2D face images, 3D face fitting results, 3D face shapes, self-occluded UV maps, UV completion results by UV-GAN, 3D face synthesis of five views, and ground truth of the frontal/profile faces. The proposed UV-GAN can generate realistic, coherent and identity-preserved UV maps for in-the-wild profile and frontal faces.

ing data, we generate non-frontal views, which preserve accurate facial details that are critical for identity discrimination. For non-frontal faces in the training data, we only increase the pose angle to generate non-frontal faces with larger self-occluded area.

VGG2 This experiment aims at assessing the impact of pose variations (frontal, three-quarter and profile views) on template matching. In [6], six templates of each subject are divided into two sets, A and B, so that each set contains templates of three different poses. And a 3×3 similarity matrix is constructed between the two sets. Each template is represented by a single vector computed by averaging the face descriptors of each face in the set. The similarity between templates is then computed as the cosine distance between the vectors representing each template.

In Tab. 3, we show the similarity matrix averaged over 368 subjects. Each row corresponds to the three pose templates from A versus the three pose templates from B. We find that (1) the similarity drops greatly when probing pairs of differing poses, *e.g.*, front-to-three-quarter and front-to-profile, implying that recognition across poses is a challenging problem; (2) recognition results are better when matching pairs with smaller pose gaps, which means that minimising pose gaps during testing, *e.g.*, face frontalisation or frontal face rotation, is highly beneficial; (3) recognition results are better when the training data contain large pose variations (*e.g.* VGG2). That is, pose-invariant feature embedding can be directly learned from such data without any special tricks; (4) by simply increasing the number of identities (*e.g.* using MS1M [14]) without having large pose variations in the training data, it is not possible to learn a pose-invariant embedding; (5) training on the proposed CASIA-augUV dataset is more beneficial for learning a pose-invariant embedding than training on the CA-

Method	Frontal-Frontal	Frontal-Profile
Human	96.24 ± 0.67	94.57 ± 1.10
Sengupta *et al*. [33]	96.40 ± 0.69	84.91 ± 1.82
Sankarana *et al*. [31]	96.93 ± 0.61	89.17 ± 2.35
Chen *et al*. [7]	98.67 ± 0.36	91.97 ± 1.70
DR-GAN [36]	97.84 ± 0.79	93.41 ± 1.17
DR-GAN+ [37]	97.84 ± 0.79	93.41 ± 1.17
Peng *et al*. [28]	98.67	93.76
CASIA-center	98.34 ± 0.44	87.77 ± 2.39
CASIA-Sphere	98.64 ± 0.24	84.39 ± 2.59
CASIA-sm	98.59 ± 0.21	87.74 ± 1.07
CASIA-sm-aug1	98.25 ± 0.42	90.14 ± 1.53
CASIA-sm-augUV	98.83 ± 0.27	93.09 ± 1.72
-Profile2Frontal	-	93.55 ± 1.67
-Frontal2Profile	-	93.72 ± 1.59
-Template2Template	-	**94.05**± 1.73
VGG2-sm	99.17 ± 0.11	93.40 ± 1.64
MS1M-sm	**99.59** ± 0.13	87.11 ± 1.47
Shape Only	67.49 ± 2.04	62.26 ± 2.57

Table 4. Verification accuracy(%) comparison on CFP dataset.

SIA dataset; (6) the performance of a model trained on the CASIA-augUV approaches the performance of a model trained in VGG2 dataset, which is 6 times larger than CASIA. Hence, it is evident that by incorporating larger pose variations in the training set using the proposed UV-GAN improves the performance.

CFP The Celebrities in Frontal-Profile (CFP) dataset [33] focuses on extreme pose face verification. The reported human performance is 96.24% on the frontal-frontal protocol and 94.57% on the frontal-profile protocol, which shows the challenge in recognising profile views.

In Tab. 4, we compare various training strategies with state-of-the-art methods. For the frontal-frontal protocol, the model MS1M-sm obtains the best verification accuracy with 99.59%, and the model VGG2-sm ranks second with an accuracy of 99.17%, which indicates that MS1M and VGG2 contain larger variations than CASIA dataset. Nevertheless, it is evident that our data augmentation on CASIA helps to improve the performance from 98.59% to 98.83%. By contrast, data augmentation by [24] (CASIA-sm-aug1) only gives an accuracy of 98.25%. The method in [24] augments the data by introducing perturbation on the parameters of the 3D shape that correspond to identity. We believe that this is not an optimal strategy, since these parameters contain important information regarding identity. To support our claim, the last row of Tab. 4 shows the performance of the 157-d identity parameters from our 3D fitting results. The recognition accuracy is 67.49% which demonstrates that even though the shape is quite low-dimensional and only estimated by 3DMM fitting it indeed contains information regarding identity.

For the frontal-profile protocol on the CFP dataset, pose augmentation during training is very effective in learning pose-invariant feature embedding. Our augmentation improves the accuracy from 87.74% (CASIA-sm) to 93.09% (CASIA-sm-augUV) and even approaches the performance of the VGG2-sm (93.40%), which is trained on a much larger dataset than CASIA (see Tab. 4). To further improve the performance, we use our UV-GAN to synthesise frontal faces from profile faces and vice versa to producing matching pairs in the testing set. We present some illustrations of the completed UV maps for profile and frontal faces in Fig. 7. From the completed UV maps and 3D face shapes, we can synthesise the face in arbitrary poses, *e.g*. frontal and profile faces.

By synthesising frontal faces from profile faces (Profile2Frontal in Tab. 4 and Fig. 7(a)) during testing and matching them, the accuracy improves by 0.46%. Similarly, synthesising profile faces from frontal faces (Frontal2Profile in Tab. 4 and Fig. 7(b)) leads to slightly better results, an improvement of 0.63%. Since our UV-GAN can generate faces with arbitrary poses from any given face, we can easily translate the pair-wise face verification problem into a more robust template verification problem. In our experiments, we synthesise frontal faces from profile faces and profile faces from frontal faces at the same time, which are denoted by Template2Template in Tab. 4. We have used a view interpolation of 15° to generate two templates. After that, we used the generated template feature centres to conduct verification, where the accuracy attains a value of 94.05%, 0.29% higher than the state-of-art method proposed by Peng *et al*. [28].

5. Conclusions

In this paper, an understudied computer vision problem, that of completion of facial UV maps that have been produced by fitting 3D face models in images. To this end, we collected a large-scale dataset of completed facial UV maps. Then, we employ global and local adversarial networks to learn identity-preserved UV completion. When we attach the completed UV map to the fitted 3D mesh, we can get faces with arbitrary poses, which can increase pose variations during training of a face recognition model and decrease pose discrepancy during testing, which lead to better performance. Experiments on both controlled and in-the-wild UV datasets confirm the effectiveness of the proposed UV completion method. The proposed method also obtains state-of-the-art verification accuracy under the CFP frontal-profile evaluation protocol.

6. Acknowledgement

J. Deng is supported by the President's Scholarship of Imperial College London. The work of S. Cheng is funded by the EPSRC project EP/N007743/1 (FACER2VM). We thank the NVIDIA Corporation for the GPU donations.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*, 2016. 5

[2] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa. The do's and don'ts for cnn-based face verification. *arXiv:1705.07426*, 2017. 5

[3] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models in-the-wild. In *CVPR*, 2017. 1, 3

[4] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016. 3

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

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *arXiv:1710.08092*, 2017. 5, 7

[7] J.-C. Chen, J. Zheng, V. M. Patel, and R. Chellappa. Fisher vector encoded deep convolutional features for unconstrained face verification. In *ICIP*, 2016. 8

[8] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv:1801.07698*, 2018. 5

[9] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *arXiv:1708.06023*, 2017. 3

[10] J. Deng, Y. Zhou, S. Cheng, and S. Zafeiriou. Cascade multi-view hourglass model for robust 3d face alignment. In *FG*, 2018. 3

[11] J. Deng, Y. Zhou, and S. Zafeiriou. Marginal loss for deep face recognition. In *CVPR Workshop*, 2017. 5

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 3

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 5

[14] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 5, 7

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 7

[17] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017. 2, 3, 4

[18] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *TOG*, 36(4):107, 2017. 2

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv:1611.07004*, 2016. 3, 4

[20] L. A. Jeni, S. Tulyakov, L. Yin, N. Sebe, and J. F. Cohn. The first 3d face alignment in the wild (3dfaw) challenge. In *ECCV*, 2016. 1

[21] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. Inversefacenet: Deep single-shot inverse face rendering from a single image. *arXiv:1703.10956*, 2017. 1

[22] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *CVPR*, 2017. 2, 4, 5

[23] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 6

[24] I. Masi, A. T. Trn, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. 2, 6, 8

[25] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv:1601.06759*, 2016. 2

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 1, 2

[27] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2, 5, 6

[28] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017. 2, 8

[29] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *TOG*, 2003. 5

[30] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. *arXiv:1611.05053*, 2016. 1

[31] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *BTAS*, 2016. 8

[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 2

[33] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016. 2, 5, 8

[34] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014. 1, 2

[35] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 2

[36] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 2, 5, 8

[37] L. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *TPAMI*, 2017. 2, 3, 4, 5, 8

[38] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *NIPS*, 2016. 2

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[40] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 4, 6

[41] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *ICCV*, 2017. 2

[42] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 2, 3, 4, 5

[43] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 2

[44] S. Zafeiriou, G. G. Chrysos, A. Roussos, E. Ververas, J. Deng, and G. Trigeorgis. The 3d menpo facial landmark tracking challenge. In *ICCV Workshop*, pages 2503–2511, 2017. 1, 3

[45] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016. 6

[46] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NIPS*, 2017. 2

[47] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 1