Deep 3D morphable model refinement via progressive growing of conditional Generative Adversarial Networks

Published Version:
Deep 3D morphable model refinement via progressive growing of conditional Generative Adversarial Networks / Galteri L.; Ferrari C.; Lisanti G.; Berretti S.; Del Bimbo A.. - In: COMPUTER VISION AND IMAGE UNDERSTANDING. - ISSN 1077-3142. - ELETTRONICO. - 185:(2019), pp. 31-42. [10.1016/j.cviu.2019.05.002]

Availability:
This version is available at: https://hdl.handle.net/11585/697538 since: 2019-09-03

Published:
DOI: http://doi.org/10.1016/j.cviu.2019.05.002

Accepted Manuscript

Deep 3D morphable model refinement via progressive growing of conditional generative adversarial networks

Leonardo Galteri, Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, Alberto Del Bimbo

PII: S1077-3142(19)30077-3
DOI: https://doi.org/10.1016/j.cviu.2019.05.002
Reference: YCVIU 2775

To appear in: *Computer Vision and Image Understanding*

Received date : 30 June 2018
Revised date : 6 May 2019
Accepted date : 12 May 2019

Please cite this article as: L. Galteri, C. Ferrari, G. Lisanti et al., Deep 3D morphable model refinement via progressive growing of conditional generative adversarial networks. *Computer Vision and Image Understanding* (2019), https://doi.org/10.1016/j.cviu.2019.05.002

*Highlights (for review)

Research Highlights

- A solution for reconstructing a fine-grained realistic 3D face model
- 3D face refinement by progressive growing of an encoder-decoder Conditional GAN
- Successfully application of progressive growing to Conditional GAN training
- Shape refinement independent from the 3D coarse reconstruction method
- Conditional GAN training on a relatively small set of examples

Deep 3D Morphable Model Refinement via Progressive Growing of Conditional Generative Adversarial Networks

Leonardo Galteri[a,**], Claudio Ferrari[a,**], Giuseppe Lisanti[b], Stefano Berretti[a], Alberto Del Bimbo[a]

[a] *University of Florence, Media Integration and Communication Center, viale Morgagni 65, Florence 50134, Italy*
[b] *University of Bologna, Department of Computer Science and Engineering, Mura Anteo Zamboni 7, Bologna 40126, Italy*

ABSTRACT

3D face reconstruction from a single 2D image is a fundamental Computer Vision problem of extraordinary difficulty. Statistical modeling techniques, such as the 3D Morphable Model (3DMM), have been widely exploited because of their capability of reconstructing a plausible model grounding on the prior knowledge of the facial shape. However, most of these techniques derive an approximated and smooth reconstruction of the face, without accounting for fine-grained details. In this work, we propose an approach based on a Conditional Generative Adversarial Network (CGAN) for refining the coarse reconstruction provided by a 3DMM. The latter is represented as a three channels image, where the pixel intensities represent the depth, curvature and elevation values of the 3D vertices. The architecture is an encoder-decoder, which is trained progressively, starting from the lower-resolution layers; this technique allows a more stable training, which leads to the generation of high quality outputs even when high-resolution images are fed during the training. Experimental results show that our method is able to produce reconstructions with fine-grained realistic details and lower reconstruction errors with respect to the 3DMM. A cross-dataset evaluation also shows that the network retains good generalization capabilities. Finally, comparison with state-of-the-art solutions evidence competitive performance, with comparable or lower error in most of the cases, and a clear improvement in the quality of the generated models.

1. Introduction

In recent years, technologies for acquiring 3D data have made substantial progress with many devices that can capture clouds of points or depth maps either statically or dynamically. Such 3D data demonstrated to be beneficial in a variety of applications, where they provide invariance to view and illumination conditions (Ioannidou et al., 2017). In particular, face and facial expression recognition constitute two application contexts where 3D data have been employed successfully, helping to improve robustness to occlusions and variations of expression, illumination and pose (Sandbach et al., 2012; Soltanpour et al., 2017). However, the diffusion and applicability of such 3D acquisition devices for face analysis is still limited: on the one hand, high-resolution scanners are typically slow and require user cooperation; on the other, depth cameras that can operate at high frame rate and without user cooperation produce low-resolution data. In both the cases, operational constraints limit the applicability of these acquisition modalities to indoor environments and close fields of view (Berretti et al., 2018). There are also multi-view stereo rigs, that represent the most commonly applied face reconstruction solution in the industry. Multi-view stereo is capable of producing 3D reconstructions with very high resolution at high capture frame rates[1]. However, these methods are based on specialized settings that include dedicated rooms and the combined views of many calibrated cameras, thus making their application more oriented to computer graphics and virtual reality for cinematographic industry.

** These authors contributed equally to this work.
Corresponding author: Tel.: +39 055 275-1394; fax: +39 055 275-1396;
e-mail: `leonardo.galteri@unifi.it` (Leonardo Galteri)

[1] See for example: http://ir-ltd.net/ or http://ten24.info/

Following a different perspective, the idea of deriving 3D information from 2D images using computer vision techniques is a research topic with a quite long tradition that dates back to '80. Now, remaining the 3D acquisition limited to certain constrained domain, the deployment of powerful machine learning tools has pushed forward this research area, with innovative and effective solutions appeared recently. Aiming to estimate the 3D geometry from single or multiple images under the most general conditions, where no *a priori* knowledge is available about the imaged scene and the capturing conditions is a very challenging task. Hence, to make the problem solvable to some extent, priors are usually assumed. In the case a 3D model of the face is reconstructed, the prior knowledge can be in the form of camera parameters and reflectance properties of the face considering either a single image, as in the shape from shading (SfS) solution (Horn and Brooks, 1989), or multiple images with different illuminations in the photometric stereo approach (Woodham, 1980). Though quite accurate reconstructions can be obtained with these solutions (Kemelmacher-Shlizerman and Basri, 2011), the given assumptions are rarely verified in real contexts. Other methods use a 3D Morphable Model (3DMM) of the face as shape prior. This statistical model limits the shape of the reconstructed face to the combination, according to a set of parameters, of an average face model and some deformation components. Different solutions have been proposed in the literature for solving for these parameters. In the original 3DMM, as firstly proposed in (Blanz and Vetter, 1999), this was formulated as the computationally onerous problem of iteratively minimizing the difference between the 2D target image and the image rendered from the 3D reconstructed model. Later works (Blanz et al., 2004; Ferrari et al., 2017b) proposed to learn the parameters via linear regression from the position of corresponding 2D and 3D landmarks. These latter solutions, though efficient, often result in coarse reconstructions that can be sensitive to inaccurate landmarks detection in the 2D images. Despite these drawbacks, the 3DMM has been the founding idea of several recent solutions that use deep neural networks to learn complex non-linear regressor functions, mapping a 2D facial image to the optimal 3DMM parameters (Tran et al., 2017a; Dou et al., 2017). However, the results of such reconstructions appear still over-smoothed, lacking of fine details of the face.

A promising idea to move a step further from the above solutions is that of starting from an initial smooth estimation of the face shape, then adding local details. A work that followed this idea, while keeping general in the assumptions, has been proposed in (Tran and Liu, 2018). In that work, a *foundation shape* is generated by a deep learning based 3DMM (Tran et al., 2017a), which is then refined by adding details generated by an encoder-decoder network. This is similar to the concept of *bump-mapping* used in Computer Graphics to separate the global shape from local details (Blinn, 1978). This idea brings quite naturally to the use of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). In the current literature of deep learning solutions, GANs have proved their capability of generating synthetic image data that are hardly distinguishable from real one (Berthelot et al., 2017). Thanks to this specific prerogative, they have found successful application in tasks such as image super-resolution (Ledig et al., 2017), image enhancement (Radford et al., 2015), image restoration (Wang et al., 2017), etc.

1.1. Contribution and paper organization

Getting inspired by the above considerations, in this work we propose a coarse-to-fine approach to reconstruct a detailed 3D face model from a single image. The approach develops on the idea of first deriving a coarse 3D shape by fitting a 3DMM. Then, the coarse shape is refined using a Conditional Generative Adversarial Network (CGAN). To this end, the 3D shape is represented as a three-channel image, where the three channels are the depth, curvature and elevation values of the vertices of the model. In addition to this, we also tested a variant of our solution, where the RGB channels of the face image are also used as input of our network. The CGAN is designed following the encoder-decoder paradigm, which is trained progressively starting from the lower-resolution layers. This technique allows a more stable training compared to traditional GANs, which ultimately leads to the generation of finer detailed 3D face models. Experimental results show that our method is able to produce reconstructions with fine-grained realistic details and lower reconstruction errors with respect to the 3DMM. A cross-dataset evaluation shows that the model retains good generalization capabilities. A comparison with state-of-the-art solutions reveals that the proposed approach is highly competitive in terms of quantitative measurements, while showing an evident superiority in generating detailed and realistic reconstructions.

In summary, our contributions are as follows:

- We design an effective and efficient solution that starting from a single image of the face is capable of deriving a fine-grained realistic 3D face model reconstruction. This is obtained by an initial coarse reconstruction followed by a refinement;
- We model the 3D face refinement step as the problem of training, with progressive growing, an encoder-decoder based Conditional GAN. Differently from (Karras et al., 2017), where a classic GAN is used, to the best of our knowledge, we are the first to successfully apply the progressive growing in the training of a conditional GAN with this architecture.
- We improve the reconstruction quality by defining an alternative solution for training the conditional GAN; in particular, we compute the adversarial-loss and the discriminator-loss considering only the depth channel, while the pixel-loss is still computed on the three-channel image;
- Through an extensive experimentation, we demonstrate that the fine-grained 3D face obtained by using the proposed solution better approximates, both quantitatively and qualitatively, a realistic face independently from the technique used to generate the coarse reconstruction given as input to the network. We also show our solution generates more realistic and detailed reconstructions with respect to state-of-the-art methods.

The rest of the paper is organized as follows: in Section 2, we summarize the closely related work on 3D face reconstruction; in Section 3, we introduce the 3D Morphable Shape Model, explain how it can be fitted to an image for generating the initial coarse estimations of the face shape, and illustrate how this serves to derive training image data with depth, curvature and elevation channels; the GAN architecture we have designed and its training are detailed in Section 4; experimental results are presented in Section 5, where we evaluate the proposed method both quantitatively, in terms of face reconstruction, and qualitatively by looking to the shape of the resulting models, also in comparison to state-of-the-art solutions; finally, conclusions and future research directions are sketched in Section 6.

2. Related work

In the general case, reconstructing a 3D face model from 2D images is extremely challenging so that most of the existing solutions rely on some assumptions in the form of prior knowledge. Keeping aside methods that do not resort to any problem simplification, and that thus result in poor reconstructions, in the following, we organize and discuss previous work into two categories according to the different priors they use.

2.1. 3D face reconstruction under constrained conditions

In the first category, there are methods that make quite strong assumptions on the data and viewing conditions, and exploit them to derive fine details in the reconstructed shape. These methods date back to '80s with the seminal works on *photometric stereo* (Woodham, 1980) and SfS (Horn, 1970; Horn and Brooks, 1989). While in photometric stereo 3D face models are reconstructed from large photo collections (Kemelmacher-Shlizerman and Seitz, 2011; Roth et al., 2015; Liang et al., 2016), the special case of SfS aims to reconstruct the face when just a single image is known (Dovgard and Basri, 2004). In both the cases, additional prior information in the form of one or more 3D models (Roth et al., 2016; Zeng et al., 2017), or statistical shape models of the face, like 3DMM (Dovgard and Basri, 2004), have been used to support the reconstruction. Though these methods show accurate and often detailed reconstructions, this is obtained at the cost of making hypothesis on the light sources and the reflectance properties of the face. Since such assumptions do not hold in practice in most of the cases, the application of these methods is limited to scenes with controlled settings.

2.2. 3D face reconstruction with shape priors

In the second category fall methods that keep general the assumptions and use priors in the form of a prototypical face model, thus reconstructing smooth shapes that, however, lack of fine details (see (A) below). An emerging trend in this category of methods is that of defining solutions that are both general and accurate. In most of the cases, this is obtained by applying a refinement step that adds details to an initially reconstructed coarse shape; deep learning solutions are mostly used for this second step (see (B)).

A. Coarse face reconstruction from shape prior – Some of the earliest methods (Vetter and Blanz, 1998; Hassner and Basri, 2006) and also more recent methods (Hassner, 2013) in this category used 3D reference models to modify the shape estimated from an input face image. For example, in (Hassner, 2013) a data-driven method was presented for estimating the 3D shape of faces viewed in single "in-the-wild" photos, where an optimization process was used to jointly maximize the similarity of appearances and depths to those of a reference model. These methods favor robustness to challenging viewing conditions over detailed reconstructions, thus they were only used to synthesize new views from unseen poses for face recognition.

The most widely recognized examples in this category are the 3DMM based fitting methods, as originally proposed in (Blanz and Vetter, 1999), and subsequently refined in other works (Romdhani and Vetter, 2003). Also these methods emphasized more the appeal of rendered face images, rather than the quantitative evaluation of the accuracy of the reconstructed face shape. Among the 3DMM variants, the most successful was proposed in (Paysan et al., 2009) that improved the 3DMM into the Basel Face Model with higher shape and texture accuracy and less correspondence artifacts. In (Booth et al., 2017b,a) an in-the-wild 3DMM was proposed by combining a statistical model of facial shape, which describes both identity and expression, with an in-the-wild texture model.

Some other reconstruction techniques fit the 3DMM surface to detected facial landmarks rather than to face intensities directly. These methods include solutions designed for videos, like in (Saito et al., 2016; Huber et al., 2016), and the CNN based approaches of (Jourabloo and Liu, 2016; Zhu et al., 2016). For example, in (Jourabloo and Liu, 2016) a face alignment method for large-pose face images was proposed that combines the powerful cascaded CNN regressor method and the 3DMM. In particular, the face alignment is formulated as a 3DMM fitting problem, where the camera projection matrix and the 3D shape parameters are estimated by a cascade of CNN-based regressors. The dense 3D shape allows designing pose-invariant appearance features for effective CNN learning. The face recognition method in (Taigman et al., 2014) also used 3D modeling of the face based on fiducial points to warp a detected facial crop to a 3D frontal mode. These latter methods, however, focus more on landmark detection and alignment than 3D shape estimation, and so do not attempt to produce detailed and discriminative facial geometries.

B. Deep face shape estimation – Recently, deep neural networks (DNN) have been applied also to the face shape estimation problem. When using deep learning approaches to reconstruct 3D faces, one main obstacle to overcome is the lack of sufficiently large amount of training data. One idea is that of generating such face shapes synthetically using a 3DMM. Following this approach, in (Richardson et al., 2016) a rather shallow network is trained on synthetic shapes with an iterative process, and facial details are also added by training an end-to-end system to additionally estimate SfS. In (Richardson et al., 2017) an end-to-end CNN framework is introduced, which derives the shape in a coarse-to-fine fashion. This architecture is composed of a network that recovers the coarse facial geometry

(CoarseNet), followed by a CNN that refines the facial features of that geometry (FineNet). Also in this case the solution space is modeled by a 3DMM.

Other methods in this category, used deep networks by emphasizing more the aspect of estimating 3D shapes from unconstrained photos (Tran et al., 2017a; Dou et al., 2017; Jackson et al., 2017; Sengupta et al., 2017). These methods estimate shapes that are highly invariant to viewing conditions, but provide only coarse 3D details. In (Tran et al., 2017a), authors proposed to use a very deep CNN to regress 3DMM parameters and facial details directly from image intensities, rather than by using the analysis by synthesis approach of earlier methods. Different from other works that reconstruct and refine the 3D face in an iterative manner using both an RGB image and an initial 3D facial shape rendering, in (Dou et al., 2017) an end-to-end DNN model was proposed that avoids the complicated 3D rendering process. In doing so, two components are integrated in the DNN architecture: a multi-task loss function, and a fusion-CNN to improve facial expression reconstruction. With the multi-task loss function, 3D face reconstruction is divided into neutral 3D facial shape reconstruction and expressive 3D facial shape reconstruction. With the fusion-CNN, features from different intermediate layers are fused and transformed for predicting the 3D expressive facial shape. In (Jackson et al., 2017), regression of a volumetric representation of the 3D face geometry from a single 2D image is directly performed using a simple CNN architecture denoted as Volumetric Regression Network that was based on the "hourglass network" (Newell et al., 2016). In (Sengupta et al., 2017), the SfSNet designs an end-to-end learning framework, which reflects a physical Lambertian rendering model for producing decomposition of an unconstrained image of a human face into shape, reflectance and illuminance. To allow for detailed reconstructions in (Sela et al., 2017) the face shape is directly estimated using a depth map. An image-to-image translation network is proposed that jointly maps the input image to a depth image and a facial correspondence map. This explicit pixel-based mapping can then be utilized to provide high-quality reconstructions of diverse faces under extreme expressions, using a purely geometric refinement process. The approach proposed in (Tewari et al., 2017) reconstructed a 3D face from a single in-the-wild color image by combining a convolutional encoder network with an expert-designed generative model that serves as decoder. The method designed in (Tran and Liu, 2018) provides detailed 3D reconstructions of faces viewed under out of plane rotations, and occlusions. Motivated by the concept of bump mapping, a layered approach is proposed, which decouples estimation of a global shape from its mid-level details (*e.g.*, wrinkles). A coarse 3D face shape is first estimated, which acts as a foundation, and then details represented by a bump map (Blinn, 1978) are layered on this foundation. A deep convolutional encoder-decoder was used to estimate such bump maps. The solution proposed in (Feng et al., 2017) exploited the Epipolar Plane Images (EPI) obtained from light-field cameras and learned CNN models that recover horizontal and vertical 3D facial curves from the respective horizontal and vertical EPIs. A 3D face reconstruction network (FaceLFnet) comprises a densely connected architecture to learn 3D facial curves from low-resolution EPIs. A framework to learn a nonlinear 3DMM from a large set of unconstrained face images, without collecting 3D face scans was proposed in (Tran and Liu, 2018). Given a face image as input, a network encoder estimates the projection, shape and texture parameters. Two decoders serve as the nonlinear 3DMM to map from the shape and texture parameters to the 3D shape and texture, respectively. An analytically-differentiable rendering layer is then used to reconstruct the original input face from the projection parameters, 3D shape, and texture. The entire network is end-to-end trainable with weak supervision. However, face reconstruction is shown just for few examples, without an extensive quantitative evaluation, and the quality of the results seem more ascribable to the face texture than to its shape.

2.3. Background on GANs for image synthesis

We are not aware of methods that use GANs, either conditional or not, to generate detailed 3D models of the face starting from a raw estimation of the shape geometry. However, in designing our reconstruction solution, we leveraged on classical GAN-based methods applied to RGB images; Therefore, in the following, we refer some relevant work that used GANs for image related tasks. GANs were first proposed in (Goodfellow et al., 2014), and subsequently modified in a series of works, for improved training (Salimans et al., 2016), or extended to unsupervised learning as with the Deep Convolutional GANs (DCGANs) (Radford et al., 2015). Since their introduction, GANs have rapidly established as state-of-the-art solutions to improve the quality of generated 2D images in a variety of image synthesis tasks. In (Denton et al., 2015), a generative parametric model was introduced capable of producing high-quality samples of natural images. This approach uses a cascade of convolutional networks within a Laplacian pyramid framework to generate images in a coarse-to-fine fashion. At each level of the pyramid, a separate generative convnet model is trained using the GAN approach. In (Odena et al., 2017), a new method for the improved training of GANs for image synthesis was introduced. Several method used GANs in image-to-image translation, where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. In (Isola et al., 2017) conditional GANs are investigated as a general-purpose solution for image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This was extended in (Zhu et al., 2017a), for learning how to translate an image from a source domain to a target domain in the absence of paired examples. In the work of (Wang et al., 2018), a new method for synthesizing high-resolution photo-realistic images from semantic label maps using conditional GANs was presented. To this end a novel adversarial loss, as well as new multi-scale generator and discriminator architectures was proposed. The solution proposed in (Zhu et al., 2017b) aims to model a distribution of possible outputs in a conditional generative modeling setting. The ambiguity of the mapping is distilled in a low-dimensional latent vector, which can be randomly sampled at test time. A generator learns to map the given input, combined with this latent code, to the output. The work in (Ledig et al., 2017), presents SRGAN, a GAN for image super-resolution. This framework is capable of inferring photo-realistic natural images for 4× up-scaling factors. This is obtained by a perceptual loss function, which consists of an adversarial loss and a content loss. In (Galteri et al., 2017) a feed-forward fully convolutional residual network model trained using a generative adversarial framework is proposed for image restoration. As specific application context, GANs have been also used to synthesize face images. In (Huang et al., 2017), a Two-Pathway GAN (TP-GAN) was proposed for photo-realistic frontal view synthesis by simultaneously perceiving global structures and local details. Four landmark located patch networks are proposed to attend to local textures in addition to the commonly used global encoder-decoder network. The work in (Tran et al., 2017b) proposed Disentangled Representation learning-GAN (DR-GAN). Starting from a non-frontal face image, this model is capable of performing face frontalization for image synthesis; At the same time, the encoder-decoder structure of the generator allows DR-GAN to learn a generative and discriminative pose-invariant representation of the face. In (Lample et al., 2017), an encoder-decoder architecture is proposed, which is trained to reconstruct images by disentangling the salient information of the image and the values of attributes directly in the latent space. As a result, after training, the model can generate different realistic versions of an input image by varying the attribute values.

Though the methods above have been inspiring for our proposed solution, they are tailored for generating 2D RGB images, while we generate a three-channel image based on depth, curvature and elevation. Despite our channels are disposed according to the same grid-like structure used for RGB images, the information carried out by each image channel is not the same, thus posing new and challenging problems about how to train GANs in a robust and effective way.

3. Coarse 3D reconstruction through 3DMM

Given a face image, we first estimate its coarse 3D reconstruction exploiting the 3D Morphable Model (3DMM) technique; then, we represent the reconstructed geometry by a three channel 2D image, where the channels represent, respectively, the *depth*, *curvature* and *elevation* of the reconstructed model. A variation of such representation has been also tested, where the RGB components of the face image are used as additional channels.

3.1. 3DMM

The coarse reconstruction represents a first estimate of the 3D face model obtained from a 2D face image. To obtain these models, we employed two different solutions, which are both based on a 3DMM.

The first approach, called *Dictionary Learning* 3DMM (*DL-3DMM*) was proposed in (Ferrari et al., 2017a): it fits the 3DMM to a face image exploiting only 2D-3D facial landmark correspondences, without accounting for the texture component. This method performs a fast fitting procedure, and can

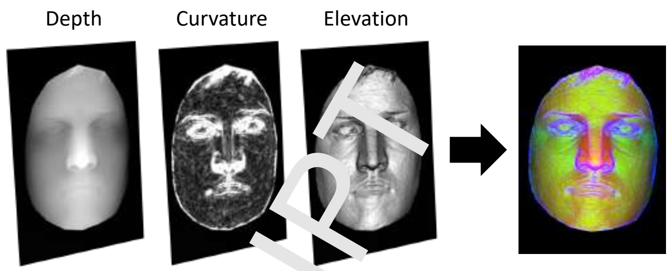

Fig. 1: Representation of the 3D face model by a three-channel image. For visualization purposes, the depth, curvature and elevation channels are shown as individual images, from left to right.

estimate the face shape fairly accurately even in the presence of strong facial expressions. The Binghamton University 3D Facial Expression dataset (BU-3DFE) (Yin et al., 2006) was used to build the average model and learn the deformation components.

The second approach is instead the one proposed by (Tran et al., 2017a): it uses the Basel Face Model (BFM) developed in (Paysan et al., 2009) and employs a deep CNN to regress the 3D shape and texture parameters of the 3DMM directly from a single RGB image, without the need of landmarks. This method is particularly robust to the subject identity, but does not model expressions. We will refer to this model as *DCNN-3DMM*.

Actually, many other 3D face modeling techniques could have fit our purposes; in fact, the proposed method aims to refine the coarse reconstruction given as input. It thus results rather independent from the coarse model that is provided, and any method can be used in practice. However, better input reconstructions will reasonably lead to more accurate refined models. Nevertheless, the above 3DMMs were chosen mainly for two reasons: (1) the first accurately reproduces facial expressions; on the opposite, (2) the second is very robust to the identity.

3.2. Facial images in depth, curvature and elevation format

The 2D representation of the 3D coarse reconstruction used in this work is inspired by the approach in (Gilani et al., 2017). Differently from the classic gray-scale depth image, this format transforms a 3D point cloud to a three-channel image. One channel contains the *depth* value of each 3D vertex; the other two contain, respectively, the *elevation* (or inclination, or polar angle) and *azimuth* values of the normal at each 3D vertex, represented in spherical coordinates.

In our case, we experimentally found that the *azimuth* channel, which encodes a geometrical property of the normal vectors as well, does not add relevant information to the final representation. On the opposite, a complementary feature for 3D meshes is the *curvature*, which encodes the degree of local variability in the surface direction. As depicted in Figure 1, in the special case of faces, the curvature highlights the shape of critical regions as the nose or eyes contour. In light of this, we decided to use the *mean curvature* (curvature in the following) instead of the azimuth property. An example of the proposed representation based on *depth*, *curvature* and *elevation* is shown in Figure 1. In order to build the elevation image, we first need to compute the normal vectors at each vertex, transform those vectors in spherical coordinates and retain the elevation (or polar angle) value. For the curvature, we applied the algorithm presented in (Cohen-Steiner and Morvan, 2003), and refer to that work for more details.

The subsequent step in the image creation is the projection of the depth, curvature and elevation values on the image plane, and rescale such values in the range $[0, 255]$ so that they can represent pixel values. This procedure must be applied consistently both for the coarsely reconstructed 3DMM and the ground-truth so that the generated images are aligned. To this aim, we estimate an orthographic projection matrix $\mathbf{P} \in \mathbb{R}^{2\times3}$ from 2D and 3D landmark correspondences. The 2D landmarks, which are detected on the RGB face images exploiting the method of (Bulat and Tzimiropoulos, 2017), are both used to fit and project the 3DMM and, independently, estimate the projection matrix for the ground-truth model so as to account for the relative difference in the models' scale. The same procedure is applied for the DCNN-3DMM; in this latter case though, the parameters to deform the 3DMM have been directly regressed from the RGB image. Thus, there is no need to fit the 3DMM, and the landmarks are only used to estimate the projection matrix to map the 3D model onto the image plane.

The projections are finally used to map the depth, curvature and elevation values on the image plane and build the three-channel images of the 3DMM and ground-truth.

Furthermore, we also experiment the use of the RGB texture as additional channels. Instead of using the original RGB face image, we perform a textured rendering of both the 3DMM and ground-truth models; the RGB texture is sampled directly from the original image and each 3D vertex is associated to a pixel value by means of the estimated projection matrix, so as to highlight the appearance changes induced by differences in the underlying geometry. This expedient also gave us the possibility to augment the training data by generating textured renderings in arbitrary 3D poses.

4. Deep generative refinement

The coarse reconstruction described in Section 3 is usually obtained as a modification of an average, smooth, model, which lacks of details. In order to obtain a fine-grained reconstruction from a single RGB face image, we propose to leverage the knowledge of several detailed 3D ground-truth models by means of a Conditional Generative Adversarial Network (CGAN). Differently from classic CGAN, the architecture is trained progressively as described in (Karras et al., 2017).

4.1. Conditional Generative Adversarial Networks (CGAN)

Conditional GANs have been specifically designed for image-to-image translation, and this makes them particularly suited for our purpose. In our solution, indeed, the generator G aims at translating the coarse reconstruction, the *condition*, to the target domain, the *ground-truth*. The discriminator D, instead, has the objective of discriminating ground-truth images from the synthetically generated ones.

Formally, the training procedure is supervised as the dataset contains paired images of the coarse model x and the correspondent detailed model y (*i.e.*, the ground-truth). The objective of conditional GANs is to learn a distribution of real detailed models given coarse input conditions as:

$$\min_G \max_D \mathbb{E}_{(x,y)} [\log D(x, y)] + \mathbb{E}_x [\log (1 - D(x, G(x)))] \ . \quad (1)$$

In our particular case, x and y are the proposed image representations of Section 3.2 for, respectively, the coarse input model (*e.g.*, the 3DMM reconstruction) and the ground truth model. The proposed solution is conditioned on x.

4.2. Progressive growing of GANs

Most of the traditional CGAN frameworks used for image-to-image translation suffer from a severe instability in the training phase (Isola et al., 2017), which is caused by high-resolution images. Indeed, networks trained with high-resolution images usually produce low quality reconstructions with unpleasant artifacts.

A solution to overcome this issue has been recently introduced in (Karras et al., 2017). This solution proposes a training procedure, which is specifically designed to cope with the problem of high-resolution image generation via GANs. The main goal of such approach is to stabilize the training algorithm so that synthetic images generated from noise would appear extremely realistic. The key idea is to start the training using very low-resolution images, then progressively increase the scale by stacking convolutional layers in the architecture. This allows the network to start learning a coarse approximation of the target distribution and consequently, as the resolution of images increases, deal with fine-grained details that affect the human perception of images.

In this framework, the generator G and the discriminator D expand their dimensions simultaneously. More specifically, after the conclusion of the training for a given resolution, the scale of the images is doubled, and a new set of convolutional layers is added to both D and G. In this way, all the existing layers maintain the learned knowledge, while remaining completely trainable for every future resolution. However, the transition between two different resolutions is not sharp and such a sudden change may significantly harm the trained weights for all the previous scales. For this reason, a transition step is introduced between the training of two resolutions. During this phase, the layers responsible for the highest level of details, *i.e.*, the last trained layers, are treated as residual blocks for which the output is the weighted sum between a $2\times$ upsampled (for the generator) or $2\times$ downsampled (for the discriminator) versions of the last resolution and the new added layer. The weighted sum is parameterized by a factor α, which is initialized to 0 and increases linearly at each iteration following a standard protocol defined in (Karras et al., 2017). In particular, α is computed as the number of the current iteration divided by the total number of iterations (*e.g.*, at iteration number five for a total of ten iterations, α is equal to 0.5).

Table 1: The structure of the discriminator and the encoder part of the generator.

G_{enc} and D			
Layer	Filter	Output shape	Params
Conv	3×3	$256 \times 256 \times 32$	864
Conv	3×3	$256 \times 256 \times 32$	9k
Conv	3×3	$256 \times 256 \times 64$	18k
MeanPool	-	$128 \times 128 \times 64$	-
Conv	3×3	$128 \times 128 \times 64$	37k
Conv	3×3	$128 \times 128 \times 128$	74
MeanPool	-	$64 \times 64 \times 128$	-
Conv	3×3	$64 \times 64 \times 128$	147k
Conv	3×3	$64 \times 64 \times 128$	147k
MeanPool	-	$32 \times 32 \times 128$	-
Conv	3×3	$32 \times 32 \times 128$	147k
Conv	3×3	$32 \times 32 \times 128$	147k
MeanPool	-	$16 \times 16 \times 128$	-
Conv	3×3	$16 \times 16 \times 128$	147k
Conv	3×3	$16 \times 16 \times 128$	147k
MeanPool	-	$8 \times 8 \times 128$	-
Conv	3×3	$8 \times 8 \times 128$	147k
Conv	3×3	$8 \times 8 \times 128$	147k
MeanPool	-	$4 \times 4 \times 128$	-
Conv	3×3	$4 \times 4 \times 128$	147k
Conv	4×4	$4 \times 4 \times 128$	262k
FC (Only D)	-	1	128
Total Parameters			**1.65M**

Table 2: The network structure of the decoder part of the generator.

G_{dec}			
Layer	Filter	Output shape	Params
Conv	4×4	$4 \times 4 \times 128$	262k
Conv	3×3	$4 \times 4 \times 128$	147k
Upsample	-	$8 \times 8 \times 128$	-
Conv	3×3	$8 \times 8 \times 128$	147k
Conv	3×3	$8 \times 8 \times 128$	147k
Upsample	-	$16 \times 16 \times 128$	-
Conv	3×3	$16 \times 16 \times 128$	147k
Conv	3×3	$16 \times 16 \times 128$	147k
Upsample	-	$32 \times 32 \times 128$	-
Conv	3×3	$32 \times 32 \times 128$	147k
Conv	3×3	$32 \times 32 \times 128$	147k
Upsample	-	$64 \times 64 \times 128$	-
Conv	3×3	$64 \times 64 \times 128$	147k
Conv	3×3	$64 \times 64 \times 128$	147k
Upsample	-	$128 \times 128 \times 128$	-
Conv	3×3	$128 \times 128 \times 64$	74k
Conv	3×3	$128 \times 128 \times 64$	37k
Upsample	-	$256 \times 256 \times 64$	-
Conv	3×3	$256 \times 256 \times 32$	18k
Conv	3×3	$256 \times 256 \times 32$	9k
Conv	3×3	$256 \times 256 \times 3$	864
Total Parameters			**1.65M**

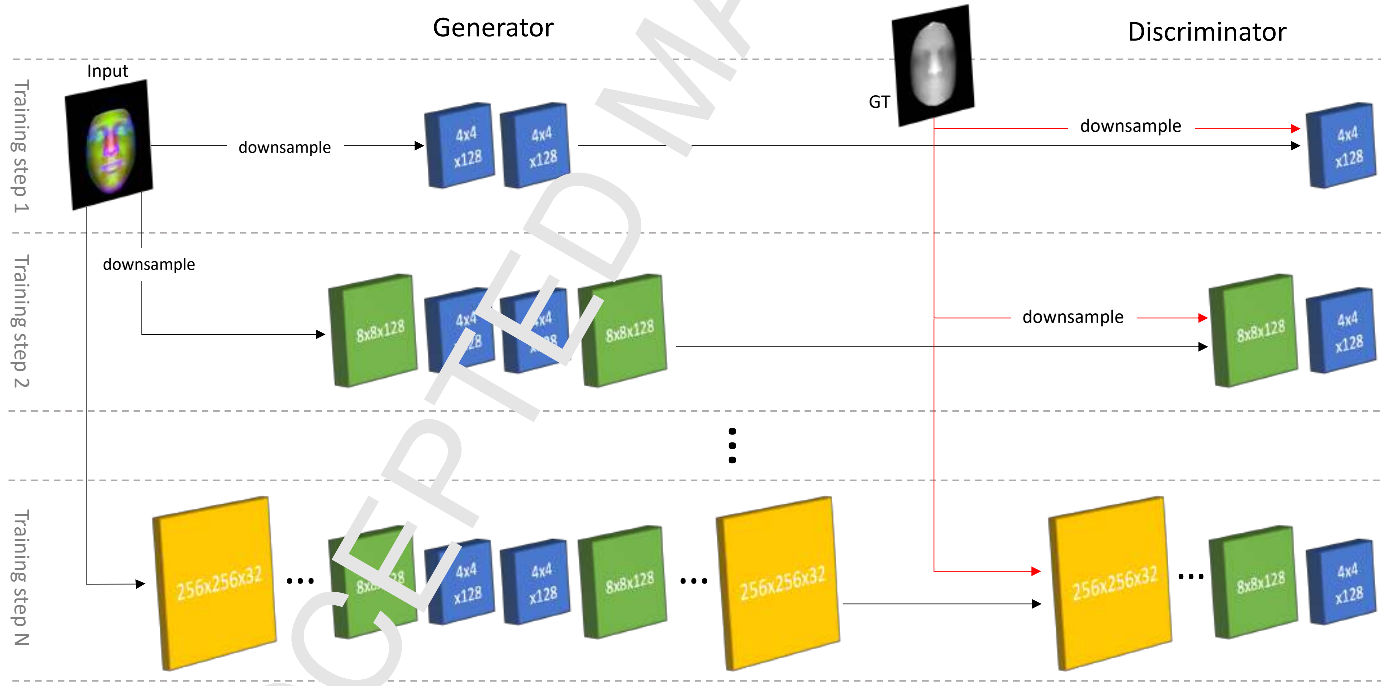

Fig. 2: Progressive refinement process. The input and output layers have been omitted for simplicity; for each training step, the number of filters for both the input and output layers is equal to the number of channels of the input image. Note that the input to the discriminator is the depth channel only.

4.3. *Progressive refinement for Conditional GANs*

We aim to exploit the benefits of progressive growth of GANs in a conditional context. For this reason, we design our generator as an encoder-decoder to transform a coarse 3DMM into a high quality detailed face model. To ensure further stability to the training of our framework, we employ the improved version of Wasserstein GAN (Gulrajani et al., 2017) as in (Karras et al., 2017). The set of weights for the discriminator are learned by minimizing the objective function:

$$\mathcal{L}_D = D(x, y) - D(x, G(x)) + \lambda(\|\nabla_{\hat{x}} D(x, \hat{x})\|_2 - 1)^2 , \quad (2)$$

where x and y are, as in Eq. (1), the proposed image representations of the coarse 3DMM and the ground truth model, respectively, and $\hat{x}$ is sampled uniformly between pairs of points belonging to the real distribution and the generator distribution.

Given the fact that our training is supervised, *i.e.*, each coarse 3DMM is paired with the relative ground-truth image, we can define the loss for the generator as a combination of two contributions:

$$\mathcal{L}_G = L_p(y, G(x)) + \kappa L_{adv}(G(x)) \ , \tag{3}$$

where

$$L_p(y, G(x)) = \|y - G(x)\|_p \ ,$$

represents the pixel loss, and

$$L_{adv}(x, G(x)) = D(x, G(x)) \ ,$$

is the adversarial loss. In this work, we use $p = 1$ in the generator loss as it has shown the best performance. We have noticed that the balance parameter, κ, has a remarkable impact on the final reconstruction. Indeed, a too low value for this parameter results in blurry outputs with missing details. This is mainly due to the fact that the adversarial component is not able to push the reconstruction towards a realistic appearance, as typical for GAN approaches. On the other hand, if κ is set too high, the reconstruction loses the required pixel-wise similarity, resulting in an output that is too different from the one of the target domain. Depending on the number of channels, C, considered in the loss computation, we empirically found that a reasonable value can be computed as $\kappa = C * 10^{-5}$.

In our particular implementation, instead of computing the adversarial loss of the discriminator on all the three input channels, *i.e.*, depth, curvature and elevation, we feed the discriminator with the depth channel only. This novel strategy grounds on two major assumptions: first, the curvature and normal vectors are two properties induced by the geometry of surfaces. If the reconstructed geometry is faithful to the ground-truth, then we expect the other two channels to be correct as well. On the other hand though, the mean curvature and the normal vectors are estimated considering local surface neighborhoods; thus, there is no guarantee that two identical curvature maps (or normal maps) are associated to the exact same depth map. As an example, a flat surface or a noisy surface, will eventually generate very similar mean curvature values on local neighborhoods. If the discriminator is trained to classify the generated images using all the channels, under the latter assumption, it could ultimately result in poor reconstructions. We instead argue that if the discriminator is trained to classify the sole depth, then the reconstructed surfaces must be as accurate as needed to be confused with the ground-truth. In any case, we also want to add a constraint on the other two features to ensure their information is exploited: we do this by imposing the pixel loss on all the three channels. As a result of this modification, the condition x, the ground-truth y and the generated image $G(x)$ in Eq. (2) and for L_{adv} in Eq. (3) represent the depth channel.

Tables 1 and 2 show the architectures for the components in our conditional GAN. Similarly to the work in (Karras et al., 2017), we employ pixel normalization after each convolutional layer of the generator; on the other hand, we do not use mini-batch standard deviation as it does not bring noticeable benefits.

We progressively train G and D starting from 4×4 downscaled images up to 256×256, expanding G in both directions simultaneously, encoder and decoder, as shown in Figure 2.

5. Experimental results

We performed a set of experiments in order to assess the validity of the proposed approach. In particular, we first show how the different elements of our solution (*i.e.*, coarse reconstruction used as input, channels considered as input for the generator, for the discriminator, and for the loss computation) influence the final performance. Then, we report on a cross dataset experiment in order to understand how well our method generalizes when considering models that come from a different distribution (*i.e.*, 3D datasets acquired with a different scanner in different conditions). Finally, we compare the proposed methods with the state-of-the-art solutions proposed by (Isola et al., 2017) and (Tran and Liu, 2018), both quantitatively and qualitatively.

5.1. Datasets

All the experiments have been carried out on two public available datasets, namely, the Face Recognition Grand Challenge dataset (FRGC) (Phillips et al., 2005) and the Bosphorus 3D Face Database (Savran et al., 2008). In particular, the FRGC dataset has been split and used both for training and for testing; whereas the Bosphorus dataset has been used only for test.

FRGC: the FRGC dataset includes 4,007 scans of 466 individuals acquired with frontal view from the shoulder level, with very small pose variations. About 60% of the faces have neutral expression, while the others show spontaneous expressions of disgust, happiness, sadness, and surprise. Scans are given as matrices of 3D points of size 480×640, with a binary mask indicating the valid points of the face (about 40K on average). 2D RGB images of the face are also available and aligned with the matrix of 3D points.

Bosphorus: the Bosphorus dataset contains 4,666 face scans of 105 subjects (60 men and 45 women, most of them of Caucasian ethnicity). Some of the scans have occlusions due to beard/moustache or short facial hair. There are about 54 face scans per subject, but 34 of these subjects have up to 31 scans due to the fewer number of expressions. On average, scans are acquired with about 30K vertices. Note that we excluded profile scans and the ones labeled as invalid.

5.2. Ground-truth model preprocessing

While the 3DMM has a fixed and clean shape, the ground-truth models are raw and may heavily differ depending on the capturing device and technique. For these reasons, they need to be preprocessed and cropped so as to eliminate surrounding areas such as ears, hairs and neck. The usual way of doing this consists in defining a sphere of fixed radius centered on the nose tip and removing the outer vertices. A drawback of this approach is that, if the sphere is too tight, the person-specific shape of the face, *e.g.*, the jawbone contour, is likely to be lost; on the contrary, if it is too large, undesired components might be included. To avoid this behavior, we considered the region defined by the intersection of the above mentioned sphere and the curve delineated by the landmarks of the facial contour. Landmarks might be provided, detected or estimated if the dataset comes with aligned pairs of RGB and range images. Finally, a median filtering was applied to remove outliers, while preserving fine-grained details.

5.3. Evaluation protocol and metric

We randomly split the FRGC individuals into three parts; the first 2/3 are used for training, for a total of 310 individuals; the remaining 1/3 of individuals and the relative models are used for test. In this way, we can ensure that an identity used for test has never been seen during the training. The models trained using the 2/3 of the FRGC are also used for the cross-dataset experiments, in which the test set is a different dataset, Bosphorus in our case.

To quantitatively evaluate our approach, we employed the *Mean Absolute Error* (MAE) measure. This is computed between the ground-truth depth image y and the estimated depth image $G(x)$ as:

$$MAE(y, G(X)) = \frac{1}{KHW} \sum_{k=1}^{K} \sum_{i=1}^{H} \sum_{j=1}^{W} \left| G(x)_{i,j}^{k} - y_{i,j}^{k} \right| \qquad (4)$$
$$\forall\, G(x)_{i,j}^{k} \neq 0 \text{ or } \forall\, y_{i,j}^{k} \neq 0\ ,$$

where K is the number of test samples, while H and W are, respectively, the height and width of the depth images.

To train and test our refinement architecture, we experimented the two solutions for 3DMM construction and fitting described in Section 3, so as to determine whether our approach can effectively generalize to coarse reconstructions obtained with different techniques and datasets.

5.4. Training settings

Data augmentation: as described in the evaluation protocol, our networks have been trained using 2/3 of the individuals in the FRGC v2.0 dataset, which results in about 2,670 depth images of 310 individuals. Unfortunately, this number is too limited for effectively training our architecture. To this end, we augmented the training data by generating novel poses as follows: given a 3D face model from the training set (coarse 3DMM and ground-truth pair), we generated a random rotation matrix $\mathbf{R}_{rand} \in \mathbb{R}^{3\times3}$, with rotation angles (yaw, pitch, roll) in the range $[\pm45, \pm20, \pm20]$, and used it to build the orthographic projection matrix $\mathbf{P}$ using a fixed 2D translation vector $\mathbf{t} \in \mathbb{R}^2$ and scale parameter matrix $\mathbf{S} \in \mathbb{R}^{2\times3}$. We then used $\mathbf{P}$ to project the pose-augmented models onto the image plane, along with the textured rendering of the input RGB image. This process is repeated 5 times for each 3D model, which results in more than $14,000$ images. During training, pixel values of each channel have been normalized in the range $[-1, 1]$. To further strengthen the procedure, we randomly crop and pad the images online during training.

Training details: the weights of the proposed architecture are initialized using a truncated normal distribution. Each resolution in our architecture has been separately trained for $10,000$ iterations with a batch size of 4 (*e.g.*, about 3 epochs with $14,000$ training samples). We train our networks using the Adam algorithm of (Kingma and Ba, 2014), with a learning rate of 10^{-5}.

5.5. Ablation Study

The proposed architecture is composed of three main components: the generator, the discriminator and the pixel loss. Each of these modules can take as input an image with different channel configurations, which can significantly change the resulting reconstruction. Thus, we conducted an ablation study so as to better assess the effect that each component has on the final reconstruction. Tests have been performed on the 1/3 of the identities of the FRGC v2.0 dataset that have not been observed in the training. Note that we use the same 3DMM technique to train and test our architecture. In the following, we comment the outcomes by referring both to the quantitative results reported in Table 3 and the qualitative examples in Figure 3. From now on, to indicate the input of each component, we will use the following naming convention for the channels: *Depth* (D), *Depth-Curvature-Elevation* (DCE), *Depth-Curvature-Elevation + RGB texture* (DCE+RGB). The network components are instead indicated with the following schema: (*Generator* | *Discriminator* | *Pixel-Loss*).

First, we performed a baseline experiment, in which the input of each module is a single-channel depth image (D | D | D). The error here is slightly lower than the coarse model, and some details are correctly generated such as the nostrils. However, the general surface is still smooth and lacking details. We then added the curvature and elevation channels as input to the generator (DCE | D | D); the latter led to an improvement in the error measures, contributing also to better reconstruct the global geometry. Adding the textured rendering to this configuration (DCE+RGB | D | D) further enhanced such effects. However, these three cases still resulted in rather smooth surfaces. In the attempt of introducing fine-grained details, we added the curvature and elevation channels as input to the discriminator and pixel loss (DCE | DCE | DCE). This solution resulted in an inconsistency between quantitative and qualitative results; we can indeed observe that error values are slightly reduced, but the corresponding 3D reconstructions are very noisy. As in the previous case, adding the RGB texture (DCE+RGB | DCE | DCE) improves the accuracy, but still, the reconstructions present a noisy surface. This behavior could be an effect of the large difference between the content of the three channels; the network, in the attempt of optimizing with respect to the three components, takes advantage of the additional geometric information provided, but introduces additional noise in the single channels. To overcome this shortcoming, we decided to remove the curvature and elevation channels from the discriminator's input, as described in Section 4. The other two network components have been instead left unchanged, leading to the last two configurations, *i.e.*, (DCE | D | DCE) and (DCE+RGB | D | DCE). Results for these configurations confirm our assumption expounded in Section 4; indeed, if the discriminator is trained to correctly classify the depth channel only, it will not be fooled unless the reconstructions are very accurate, which is actually what we aim at obtaining. The other two channels, instead, provide the right geometrical information needed to generate faithful and self-consistent surfaces. Even though these solutions do not provide the best quantitative results, it can be appreciated that the resulting 3D reconstructions take both the advantages

Fig. 3: Reconstructions with different network configurations. The sole depth channel is not sufficient to reproduce fine grained details (third to fifth column with reconstructed models); the curvature and elevation channels bring geometrical information, but induce noise to the reconstructions (sixth and seventh columns); reconstructing all the three channels while discriminating only the depth provides good results both in terms of geometry, surface details and absence of noise (rightmost two columns).

Table 3: Mean absolute error (MAE) computed on the test set of the FRGC v2.0 dataset and on the Bosphorus dataset. Input for the generator, discriminator and loss computation are indicated as: *Depth* (D); *Depth-Curvature-Elevation* (DCE); *Depth-Curvature-Elevation + RGB Texture* (DCE+RGB).

				FRGC v2.0	Bosphorus
Input model	**Generator**	**Discriminator**	**Pixel-Loss**	**MAE**	**MAE**
DL-3DMM (Ferrari et al., 2017b)	D	D	D	0.095 ± 0.030	0.176 ± 0.054
	DCE	D	D	0.079 ± 0.026	0.178 ± 0.058
	DCE	DCE	DCE	0.073 ± 0.026	0.160 ± 0.057
	DCE	D	DCE	0.078 ± 0.025	0.159 ± 0.057
	DCE + RGB	D	D	0.062 ± 0.022	0.218 ± 0.058
	DCE + RGB	DCE	DCE	0.064 ± 0.020	0.237 ± 0.058
	DCE + RGB	D	DCE	0.065 ± 0.023	0.233 ± 0.062
	Coarse			0.119 ± 0.039	0.175 ± 0.057
DCNN-3DMM (Tran et al., 2017a)	D	D	D	0.079 ± 0.025	0.105 ± 0.058
	DCE	D	D	0.066 ± 0.022	0.118 ± 0.059
	DCE	DCE	DCE	0.067 ± 0.023	0.124 ± 0.062
	DCE	D	DCE	0.069 ± 0.024	0.124 ± 0.058
	DCE + RGB	D	D	0.060 ± 0.021	0.123 ± 0.053
	DCE + RGB	DCE	DCE	0.065 ± 0.021	0.119 ± 0.052
	DCE + RGB	D	DCE	0.067 ± 0.023	0.151 ± 0.052
	Coarse			0.140 ± 0.039	0.132 ± 0.060

of the previous solutions by providing fine-grained details on a well reconstructed global shape. In any case, the refined models obtain a lower error with respect to both the coarse models.

In Figure 4, we report some absolute error heatmaps, with respect to the ground-truth (GT), for each coarse reconstruction and the respective refined models. From the figure, the lower error obtained with the refinement compared to the coarse counterpart can be appreciated. Note that our framework aims to enrich with fine-grained details an initial shape estimate; thus, it cannot completely eliminate large errors due to the coarse reconstruction, but still, it is able to reduce the error and correctly generate the details. This is evident in the example in Figure 4, bottom row. In some other cases, the refinement can fail because of wrong landmark localizations. However, note that this only implies that the network is not able to reconstruct a shape resembling the ground-truth, but is still able to improve the shape and generate fine details.

5.6. *Cross-dataset test on Bosphorus*

We performed a cross-dataset experiment in order to understand whether our architecture can generalize to new unseen data from a different distribution, *i.e.*, a different dataset. For this experiment, we considered the Bosphorus dataset as test set. Results for this test are reported in the last column of Table 3. Here, the general error is higher with respect to the one obtained on the FRGC. This result was predictable because the proposed solution is trying to reconstruct fine details on face models from a dataset that has been acquired with a different device and presents different surface characteristics. However, also in this case, the refined models obtain a lower error with respect to the coarse reconstructions. In this regard, we note that in a cross-dataset scenario, including the textured RGB renderings has the effect of increasing the error. This is not surprising since the RGB face images have large appearance differences across the datasets and thus belong to very different distributions. Nonetheless, one of the claimed advantages of the proposed approach was that it is dataset independent to a great extent; in fact, the 3-channel image representation of the coarse 3DMM will be always the same, regardless of the face images it has been fit to. This makes our approach applicable to any dataset, once the coarse reconstruction approach has been fixed. We further noted that, in this dataset, the landmark detector fails in correctly localizing the face landmarks more often with respect to FRGC, which might be a concur-

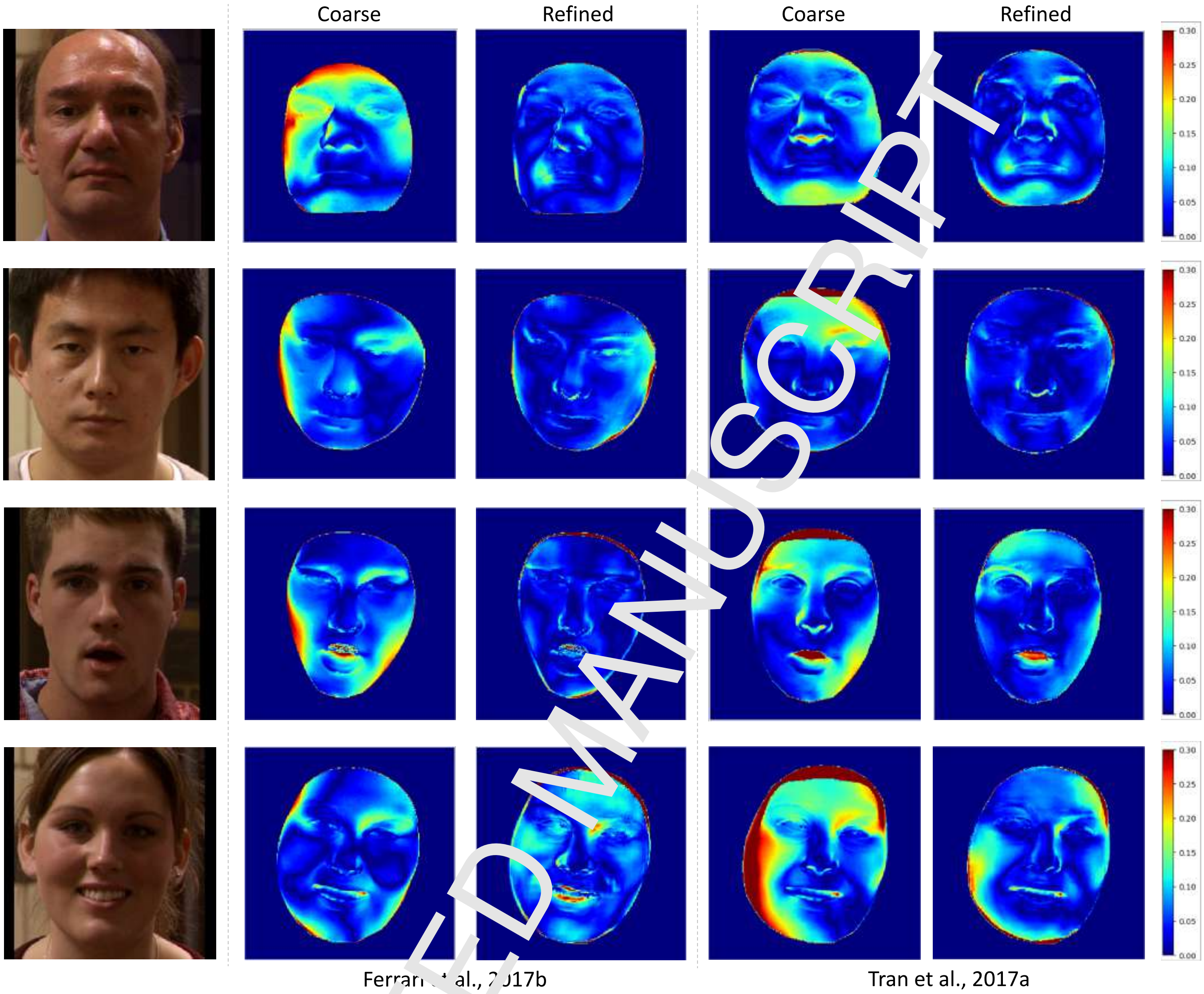

Fig. 4: Error heat maps with respect to the GT. Coarse and refined models (DCE | D | DCE) are shown in the first three rows, while the last row shows a critical case.

rent cause of the higher general error. Moreover, this dataset includes stronger expressions, pose variations and occlusions.

5.7. Comparison with state-of-the-art

In order to present a complete evaluation, we compared our method with two state-of-the-art solutions, both quantitatively and qualitatively. In particular, we trained a recent CGAN architecture, namely *Pix2Pix* (Isola et al., 2017), to refine the two coarse models considered in this work. From an architectural point of view, it adopts the U-Net as a generator, and it embodies a patch discriminator. We trained Pix2Pix on our 256 × 256 training images with the default settings for 20 epochs. We compared also against the end-to-end framework proposed in (Tran and Liu, 2018), which takes an RGB face image and produces a detailed 3D model through a 3DMM solution. An important reason that guided us in choosing this specific approach is that the coarse model generated by the first step of that method is actually the 3DMM of (Tran et al., 2017a), which is used as coarse model in this work as well. This allowed us to disentangle the two steps and derive a comparison focusing on the refinement step.

Performance of our solution with respect to the state-of-the-art are reported in Table 4, in terms of mean absolute error; results are shown for both FRGC and Bosphorus. For comparison, we selected our two best configurations as resulting from the ablation study, that are (DCE | D | DCE) and (DCE+RGB | D | DCE). For Pix2Pix instead, we chose the two configurations that takes both DCE and DCE+RGB as input and outputs the depth channel only, *i.e.*, (DCE | D | D) and (DCE+RGB | D | D). This because, as reported in Figure 3, trying to reconstruct with respect to all the three DCE channels results in a very noisy reconstructed surface; we experimentally found that such behavior applies also for Pix2Pix. Results in Table 4 show that our solution gets favorable MAE with respect to the compared approaches on the FRGC dataset; on the Bosphorus dataset, errors tend to be generally higher. As a first note, we wish to point out that results of the method in (Tran and Liu, 2018) (indicated with a '*' in the table) are not totally compa-

Table 4: Comparison with the state-of-the-art in terms of mean absolute error (MAE) computed on the test set of the FRGC v2.0 dataset and on the Bosphorus dataset. Input for the generator, discriminator and loss computation are indicated as: *Depth* (D); *Depth-Curvature-Elevation* (DCE); *RGB Texture* (RGB); *Depth-Curvature-Elevation + RGB Texture* (DCE+RGB).

					FRGC v2.0	Bosphorus
Input model	**Refinement**	**Generator**	**Discriminator**	**Pixel-Loss**	**MAE**	**MAE**
DL-3DMM (Ferrari et al., 2017b)	*Our*	DCE	D	DCE	0.078 ± 0.025	0.159 ± 0.057
	Pix2Pix	DCE	D	D	0.083 ± 0.029	0.164 ± 0.062
	Our	DCE+RGB	D	DCE	0.065 ± 0.023	0.233 ± 0.062
	Pix2Pix	DCE+RGB	D	D	0.079 ± 0.028	0.211 ± 0.062
DCNN-3DMM (Tran et al., 2017a)	*Our*	DCE	D	DCE	0.069 ± 0.024	0.124 ± 0.058
	Pix2Pix	DCE	D	D	0.067 ± 0.024	0.119 ± 0.064
	Our	DCE+RGB	D	DCE	0.067 ± 0.023	0.151 ± 0.052
	Pix2Pix	DCE+RGB	D	D	0.062 ± 0.021	0.123 ± 0.055
DCNN-3DMM (Tran et al., 2017a)	(Tran and Liu, 2018)	RGB	–	D	0.129 ± 0.037	0.130 ± 0.058*

*It was not possible to use 848 out of 3500 test images on Bosphorus

rable because the approach failed in detecting and thus reconstructing 848 faces out of 3500 of the test set of the Bosphorus dataset, most likely for the tight crop or the strong expressions portrayed. With respect to Pix2Pix, instead, in this dataset our approach seems to struggle; however, qualitative results in Figures 5 and 6 show that the reconstructed models of Pix2Pix present a pronounced noise on the whole surface in both the datasets, and the reconstructions appear evidently worse than ours. Indeed, the method is actually able to somewhat capture the underlying geometry, but fails in reproducing a pleasant and clean detailed surface. The lower error could be ascribed to the noisier nature of the Bosphorus scans (note Figure 6, last two rows); most of the ground-truth models with respect to which the error is computed, present a noise pattern that is similar to the one produced by the Pix2Pix reconstructions. Thus, we argue that when comparing the two, the final error results lower. Our solution, instead, is able to generate highly detailed and geometrically faithful reconstructions, introducing far less noise. Finally, consistently with the quantitative results reported, the variants including the RGB texture eventually led to worse reconstructions in a cross-dataset scenario.

6. Conclusions

In this work, we proposed an approach based on a Conditional Generative Adversarial Network (CGAN) for refining the coarse reconstruction of face images provided by a 3DMM. The reconstruction is represented as a three channel image, where the pixel intensities represent the depth, curvature and elevation values of the 3D model's vertices. We proposed an encoder-decoder architecture, which is trained progressively; this technique allowed a more stable training, which led to the generation of artifact-free images even at higher resolutions. Experimental results showed that our method can generate reconstructions with fine-grained realistic details for all the two different coarse 3DMM reconstructions taken into account. A cross-dataset evaluation finally showed that the architecture retains good generalization capabilities as well. However, our approach is not exempt from limitations; first, if the shape of the 3DMM differs too much with respect to the ground-truth ones, the network might eventually overfit the data in the attempt of transforming the shapes and thus lose its generalization capabilities or, on the contrary, fail in generating pleasant outputs. Another limitation is that if we want to change the coarse 3D reconstruction model to be refined, a new instance of the network has to be trained from scratch. Even though the training procedure is rather fast and does not require as many images as other architectures, we still might want to investigate if a feasible solution to make it independent from the coarse 3D input can be found.

Overall, we demonstrated that a progressive CGAN can be effectively trained on distinctive image data and employed to generate highly detailed 3D surfaces from their smoother counterparts. The solutions that have been investigated and presented in this manuscript actually represent only a small portion of the possible alternatives, for which there is a lot of room for improvements. As an example, we will further investigate how to exploit the correlations that occur between the three channels encoding surface geometric properties to our advantage.

Acknowledgements

The Titan Xp GPUs used for this research were donated by the NVIDIA Corporation.

References

Berretti, S., Daoudi, M., Turaga, P., Basu, A., 2018. Represenation, analysis and recognition of 3D humans: A survey. ACM Trans. on Multimedia Computing, Communication and Applications 14, 1–35.

Berthelot, D., Schumm, T., Metz, L., 2017. BEGAN: Boundary equilibrium generative adversarial networks. CoRR abs/1703.10717.

Blanz, V., Mehl, A., Vetter, T., Seidel, H.P., 2004. A statistical method for robust 3D surface reconstruction from sparse data, in: Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004., pp. 293–300. doi:10.1109/TDPVT.2004.1335212.

Blanz, V., Vetter, T., 1999. A morphable model for the synthesis of 3D faces, in: ACM Conf. on Computer Graphics and Interactive Techniques.

Blinn, J.F., 1978. Simulation of wrinkled surfaces, in: Annual Conference on Computer Graphics and Interactive Techniques, ACM. pp. 286–292. doi:10.1145/800248.507101.

Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S., 2017a. 3D face morphable models "in-the-wild", in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5464–5473. doi:10.1109/CVPR.2017.580.

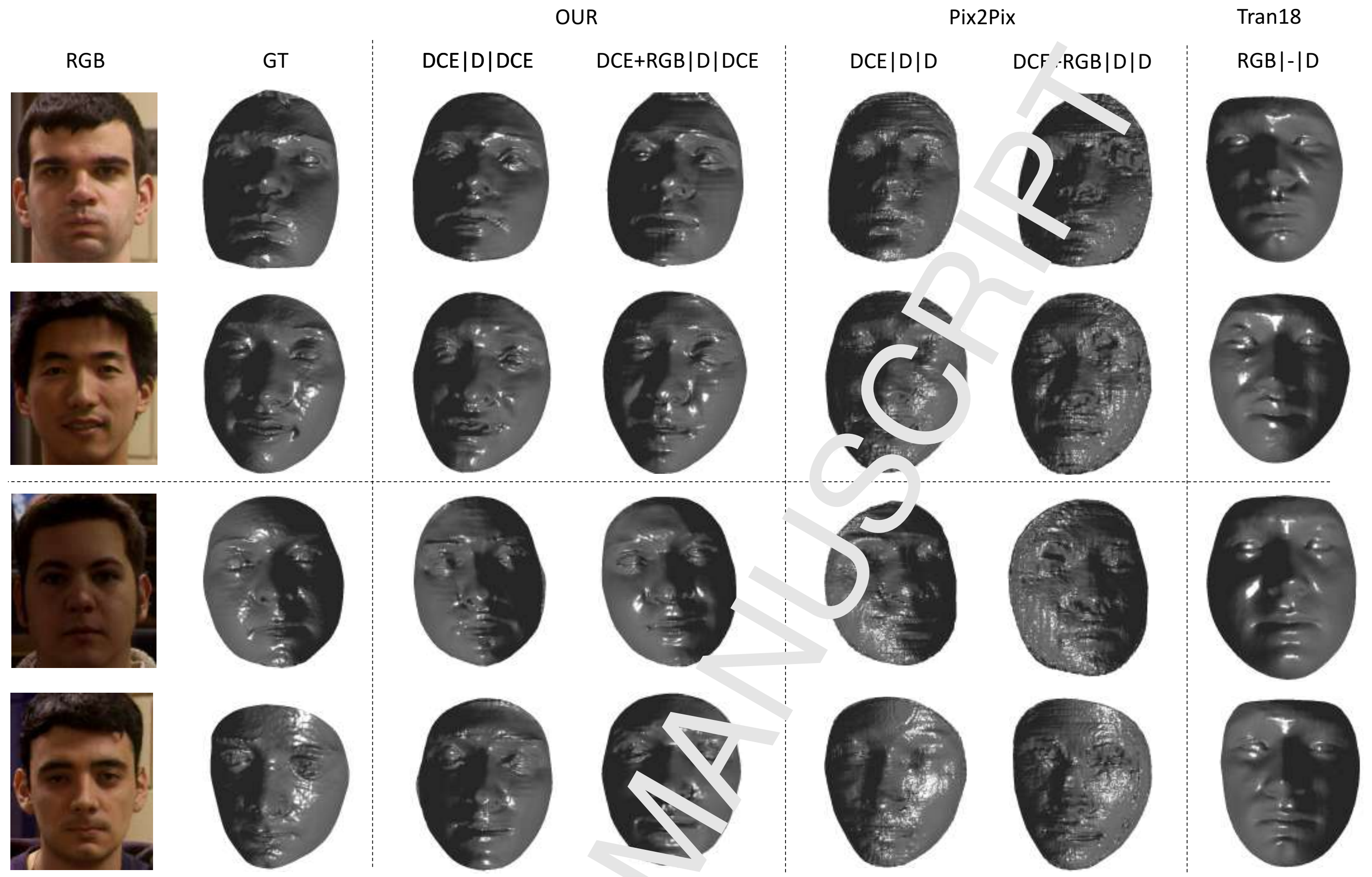

Fig. 5: Qualitative results on the FRGC test set. The first two rows refer to the DL-3DMM coarse model, while the last two rows to the DCNN-3DMM. Qualitatively, our solution generates accurate and cleaner reconstructions with respect to state-of-the-art approaches.

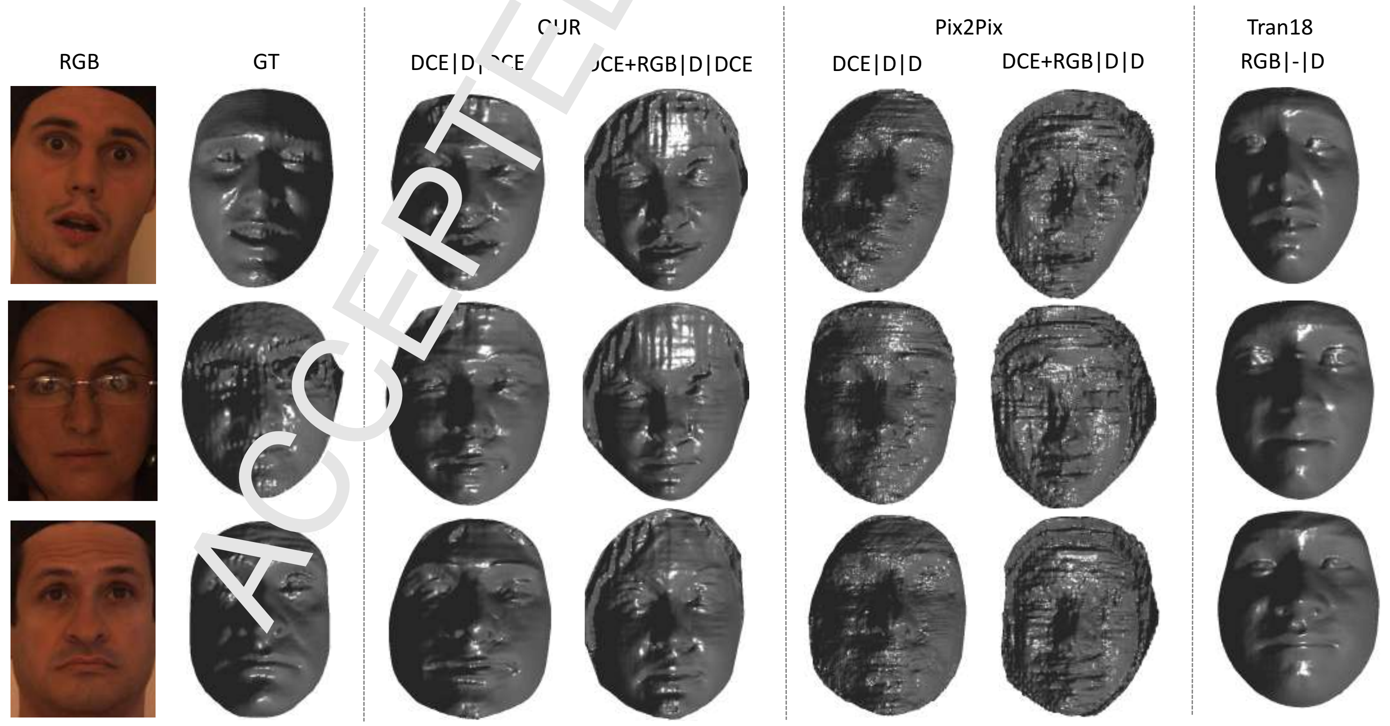

Fig. 6: Qualitative results on the Bosphorus dataset. Our approach is still able to generate pleasant outputs even for expressive faces. Much heavier noise is introduced by the Pix2Pix method. Adding the RGB texture results critical; note how the headgear in the images is interpreted as hair by the networks.

Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S., 2017b. Large scale 3D morphable models. Int. Journal of Computer Vision 126, 233–254.

Bulat, A., Tzimiropoulos, G., 2017. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks), in: Int. Conf. on Computer Vision (ICCV).

Cohen-Steiner, D., Morvan, J.M., 2003. Restricted delaunay triangulations and normal cycle, in: Annual Symposium on Computational Geometry, ACM, New York, NY, USA. pp. 312–321. doi:10.1145/777792.777839.

Denton, E.L., Chintala, S., Szlam, A., Fergus, R., 2015. Deep generative image models using a laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing Systems (NIPS), pp. 1486–1494.

Dou, P., Shah, S.K., Kakadiaris, I.A., 2017. End-to-end 3D face reconstruction with deep neural networks, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1503–1512. doi:10.1109/CVPR.2017.164.

Dovgard, R., Basri, R., 2004. Statistical symmetric shape from shading for 3D structure recovery of faces, in: European Conf. on Computer Vision (ECCV), Springer Berlin Heidelberg. pp. 99–113.

Feng, M., Gilani, S.Z., Wang, Y., Mian, A., 2017. 3D face reconstruction from light field images: A model-free approach. arXiv 1711.05953 abs/1711.05953.

Ferrari, C., Lisanti, G., Berretti, S., Del Bimbo, A., 2017a. A dictionary learning-based 3D morphable shape model. IEEE Trans. on Multimedia 19, 2666–2679. doi:10.1109/TMM.2017.2707341.

Ferrari, C., Lisanti, G., Berretti, S., Del Bimbo, A., 2017b. Investigating nuisance factors in face recognition with dcnn representation, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 583–591. doi:10.1109/CVPRW.2017.86.

Galteri, L., Seidenari, L., Bertini, M., Del Bimbo, A., 2017. Deep generative adversarial compression artifact removal, in: IEEE Int. Conf. on Computer Vision (ICCV), pp. 4836–4845. doi:10.1109/ICCV.2017.517.

Gilani, S.Z., Mian, A., Eastwood, P., 2017. Deep, dense and accurate 3D face correspondence for generating population specific deformable models. Pattern Recognition 69, 238–250. doi:https://doi.org/10.1016/j.patcog.2017.04.013.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein GANs. arXiv 1704.00028 .

Hassner, T., 2013. Viewing real-world faces in 3D, in: IEEE Int. Conf. on Computer Vision (ICCV), pp. 3607–3614. doi:10.1109/ICCV.2013.448.

Hassner, T., Basri, R., 2006. Example based 3D reconstruction from single 2D images, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW), pp. –. doi:10.1109/CVPRW.2006.76.

Horn, B., 1970. Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View. Technical Report AI-TR-232. MIT.

Horn, B.K.P., Brooks, M.J. (Eds.), 1989. Shape from Shading. MIT Press, Cambridge, MA, USA.

Huang, R., Zhang, S., Li, T., He, R., 2017. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis, in: IEEE Int. Conf. on Computer Vision (ICCV), pp. 2458–2467. doi:10.1109/ICCV.2017.267.

Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, P., Christmas, W., Ratsch, M., Kittler, J., 2016. A multiresolution 3D morphable face model and fitting framework, in: Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications.

Ioannidou, A., Chatzilari, E., Nikolopoulos, S., Kompatsiaris, I., 2017. Deep learning advances in computer vision with 3d data: A survey. ACM Computer Survey 50, 20:1–20:38. doi:10.1145/3042064.

Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976. doi:10.1109/CVPR.2017.632.

Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G., 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression, in: IEEE Int. Conf. on Computer Vision (ICCV), pp. 1031–1039.

Jourabloo, A., Liu, X., 2016. Large-pose face alignment via CNN-based dense 3D model fitting, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4188–4196. doi:10.1109/CVPR.2016.454.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of GANs for improved quality, stability, and variation. arXiv 1710.10196 .

Kemelmacher-Shlizerman, I., Basri, R., 2011. 3D face reconstruction from a single image using a single reference face shape. IEEE Trans. on Pattern Analysis and Machine Intelligence 33, 394–405. doi:10.1109/TPAMI.2010.63.

Kemelmacher-Shlizerman, I., Seitz, S.M., 2011. Face reconstruction in the wild, in: IEEE Int. Conf. on Computer Vision (ICCV), p. 17461753. doi:10.1109/ICCV.2011.6126439.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv 1412.6980 .

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., DENOYER, L., Ranzato, M.A., 2017. Fader networks:manipulating images by sliding attributes, in: Advances in Neural Information Processing Systems (NIPS), pp. 5967–5976.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 105–114. doi:10.1109/CVPR.2017.19.

Liang, S., Shapiro, L.G., Kemelmacher-Shlizerman, I., 2016. Head reconstruction from internet photos, in: European Conf. on Computer Vision (ECCV), Springer. pp. 360–374.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation, in: European Conf. on Computer Vision (ECCV), Springer International Publishing. pp. 483–499.

Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier GANs, in: Int. Conf. on Machine Learning (ICML), pp. 2642–2651.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T., 2009. A 3D face model for pose and illumination invariant face recognition, in: IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), pp. 296–301.

Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W., 2005. Overview of the face recognition grand challenge, in: IEEE Workshop Face Recognition Grand Challenge Experiments.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR abs/1511.06434.

Richardson, E., Sela, M., Kimmel, R., 2016. 3D face reconstruction by learning from synthetic data, in: IEEE Int. Conf. on 3D Vision (3DV), pp. 460–469. doi:10.1109/3DV.2016.56.

Richardson, E., Sela, M., Or-El, R., Kimmel, R., 2017. Learning detailed face reconstruction from a single image, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5553–5562. doi:10.1109/CVPR.2017.589.

Romdhani, S., Vetter, T., 2003. Efficient, robust and accurate fitting of a 3d morphable model, in: IEEE Int. Conf. on Computer Vision (ICCV), pp. 59–66. doi:10.1109/ICCV.2003.1238314.

Roth, J., Tong, Y., Liu, X., 2015. Unconstrained 3D face reconstruction, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2606–2615. doi:10.1109/CVPR.2015.7298876.

Roth, J., Tong, Y., Liu, X., 2016. Adaptive 3D face reconstruction from unconstrained photo collections, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4197–4206. doi:10.1109/CVPR.2016.455.

Saito, S., Li, T., Li, H., 2016. Real-time facial segmentation and performance capture from RGB input, in: European Conf. Computer Vision (ECCV), Springer International Publishing. pp. 244–261.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs, in: Int. Conf. on Neural Information Processing Systems (NIPS), pp. 2234–2242.

Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L., 2012. Static and dynamic 3D facial expression recognition: A comprehensive survey. Image and Vision Computing 30, 683–697.

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L., 2008. Bosphorus database for 3d face analysis, in: European Workshop on Biometrics and Identity Management, Springer. pp. 47–56.

Sela, M., Richardson, E., Kimmel, R., 2017. Unrestricted facial geometry reconstruction using image-to-image translation, in: IEEE Int. Conf. on Computer Vision (ICCV), pp. 1576–1585.

Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D., 2017. SfSNet: Learning shape, reflectance and illuminance of faces in the wild. ArXiv 1712.01261 .

Soltanpour, S., Boufama, B., Wu, Q.J., 2017. A survey of local feature methods for 3d face recognition. Pattern Recognition 72, 391 – 406. doi:https://doi.org/10.1016/j.patcog.2017.08.003.

Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. DeepFace: Closing the gap to human-level performance in face verification, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708.

Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P., Theobalt, C., 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction, in: IEEE Int. Conf. on Computer Vision Workshops (ICCVW), pp. 1274–1283.

Tran, A.T., Hassner, T., Masi, I., Medioni, G., 2017a. Regressing robust and discriminative 3D morphable models with a very deep neural network, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5163–5172.

Tran, L., Liu, X., 2018. Nonlinear 3D face morphable model, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 7346–7355.

Tran, L., Yin, X., Liu, X., 2017b. Disentangled representation learning GAN for pose-invariant face recognition, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1283–1292. doi:10.1109/CVPR.2017.141.

Vetter, T., Blanz, V., 1998. Estimating coloured 3D face models from single images: An example based approach, in: European Conf. on Computer Vision (ECCV), Springer Berlin Heidelberg. pp. 499–513.

Wang, P., Zhang, H., Patel, V.M., 2017. Generative adversarial network-based restoration of speckled sar images, in: IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 1–5. doi:10.1109/CAMSAP.2017.8313133.

Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional GANs, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Woodham, R.J., 1980. Photometric method for determining surface orientation from multiple images. Optical Engineering 19, 139–144. doi:10.1117/12.7972479.

Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M., 2006. A 3D facial expression database for facial behavior research, in: IEEE Int. Conf. on Automatic Face and Gesture Recognition.

Zeng, D., Zhao, Q., Long, S., Li, J., 2017. Examplar coherent 3D face reconstruction from forensic mugshot database. Image and Vision Computing 58, 193–203. doi:https://doi.org/10.1016/j.imavis.2016.03.001.

Zhu, J., Park, T., Isola, P., Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE Int. Conf. on Computer Vision (ICCV), pp. 2242–2251. doi:10.1109/ICCV.2017.244.

Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E., 2017b. Toward multimodal image-to-image translation, in: Advances in Neural Information Processing Systems (NIPS), pp. 465–476.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z., 2016. Face alignment across large poses: A 3D solution, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 146–155. doi:10.1109/CVPR.2016.23.