Deep Learning for 3D Face Reconstruction

Claudio Ferrari, Stefano Berretti, Alberto Del Bimbo

claudio.ferrari@unifi.it, stefano.berretti@unifi.it, alberto.delbimbo@unifi.it

https://sites.google.com/unifi.it/3dface-tutorial-cvpr20

Department of Information Engineering (DINFO) &
Media Integration and Communication Center (MICC)

University of Florence (UNIFI), Florence, Italy

Opening Remark

In this part of the tutorial, we will present some recent deep learning based works that address the face reconstruction from single images problem, and the basics of deep learning applied to 3D data;

We will focus mainly on those that employ the 3D Morphable Model in some way (the vast majority);

The goal is to give an overview of what has been done, and what can be done;

A basic knowledge of deep learning and CNNs (basic architectures, training procedures etc) is assumed.

Deep Learning for 3D Vision

The advent of deep learning techniques based on Convolutional Neural Networks (CNN) has drastically changed the way computer vision problems are being addressed;

CNNs were specifically designed to work with 2D image data;

The diffusion of such techniques in the 3D vision field has had a slower expansion because of (1) the **diverse data representation** i.e. irregular 3D data against regular RGB imagery and (2) **lack of 3D data**;

Despite this, deep learning techniques are currently being applied with promising results also to 3D data

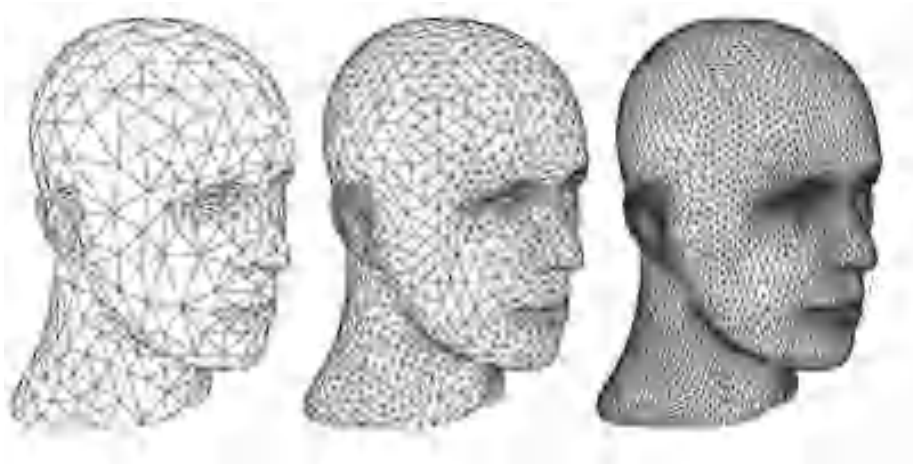
Representation Issue

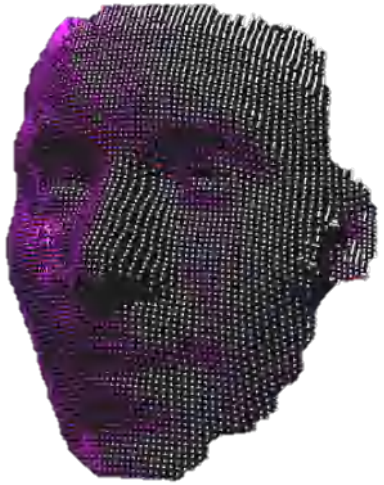
3D data can be represented in many different ways

- RGB(D) or depth images; → **Can directly apply CNN**
- Point-cloud
- Triangulated Meshes
- Volumetric

Need dedicated architectures and operators

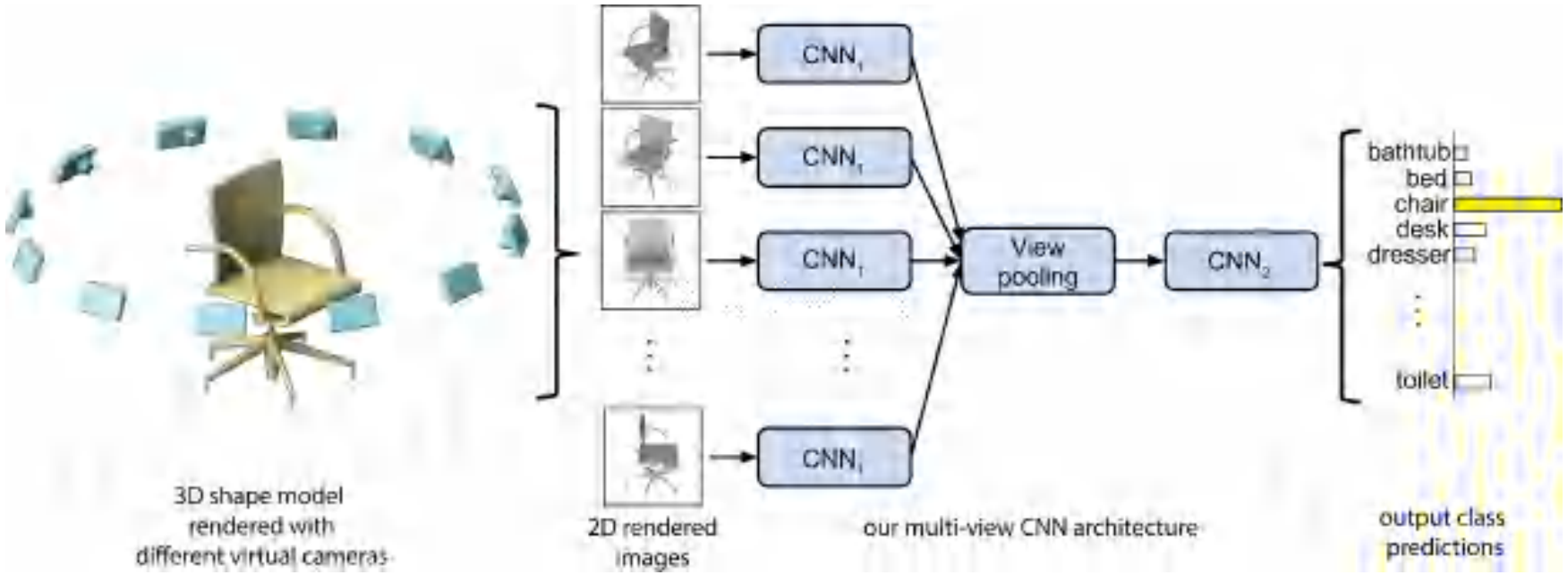
- …





A workaround

- A possible way to address the problem consists in rendering 2D images of the object with virtual cameras;
- The multiple views are then processed with usual convolutional networks.
- Not really exploiting the real 3D information though…

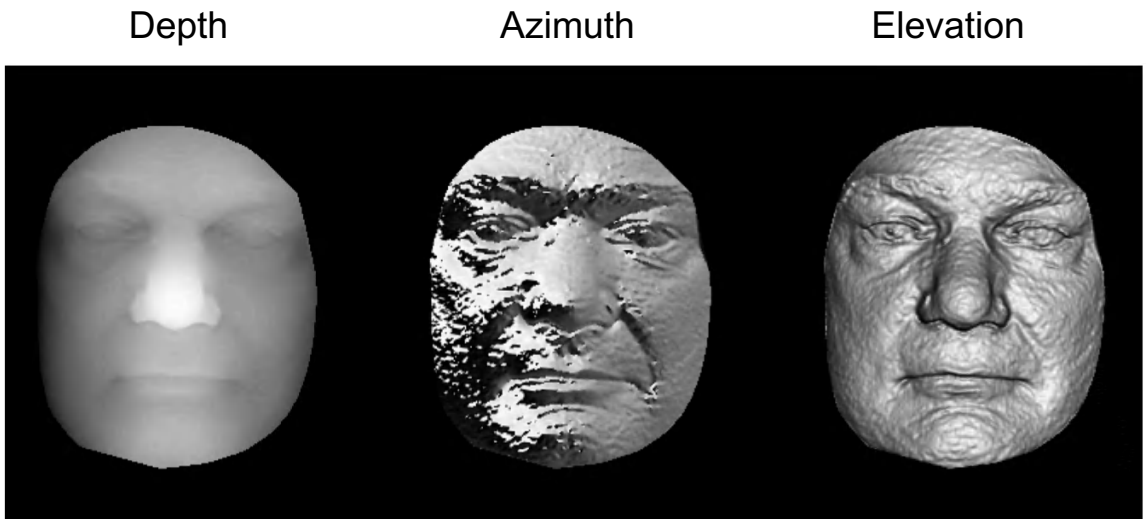


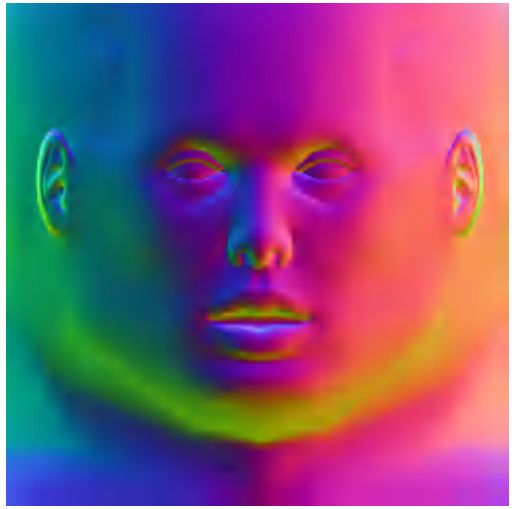
Su, Hang, et al. "Multi-view convolutional neural networks for 3d shape recognition." Proceedings of the IEEE international conference on computer vision. 2015.

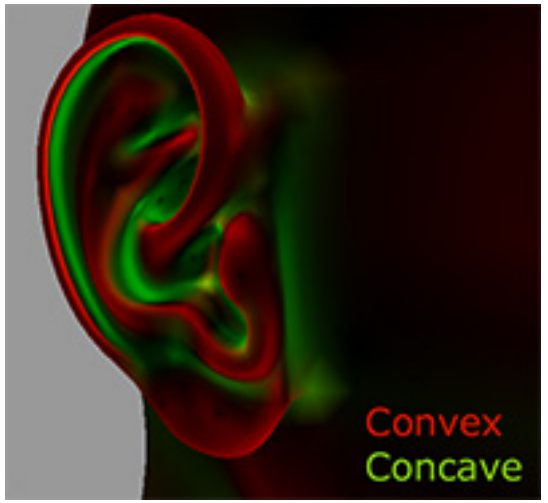
Looking on the bright side

3D data contain more information with respect to an RGB image. One way to extend the previous idea is that of computing some surface property and encode the information into an image, for example:

- Depth (z coordinate)
- Direction of the normal vectors (ex: azimuth and elevation angles)
- Curvature
- Possibly others or a composition of the above





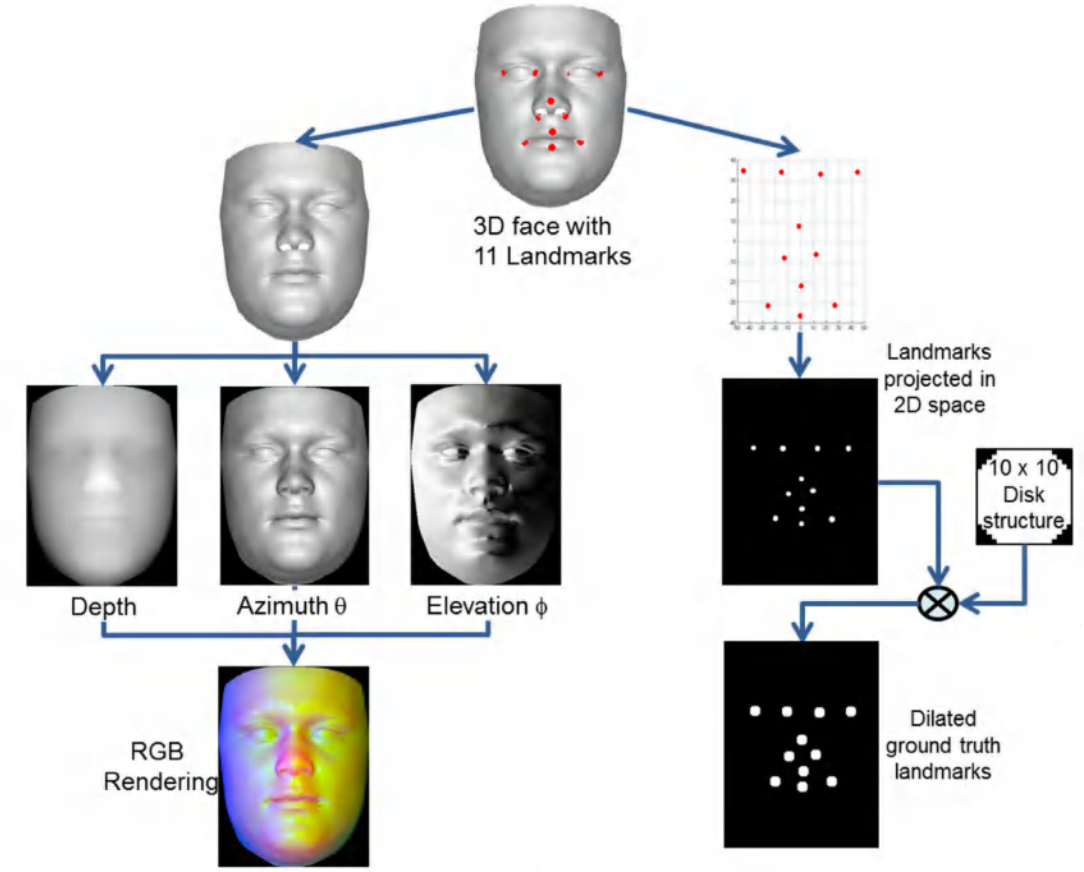


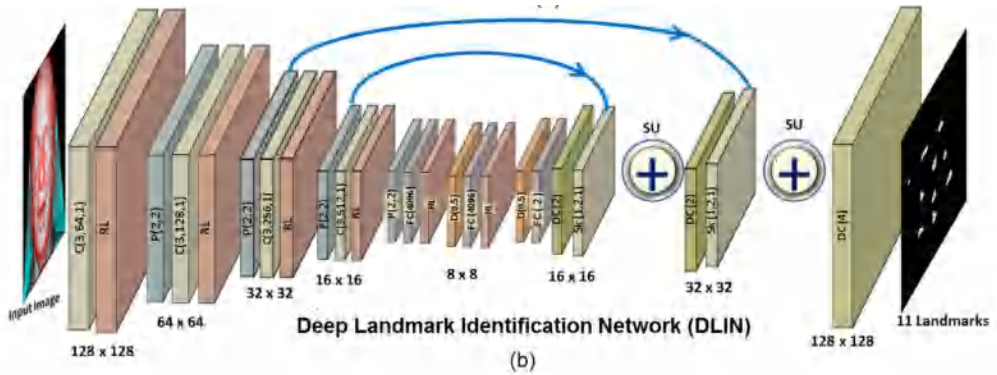
An example: 3D landmark detection

The depth, azimuth and elevation angles of the surface normal vectors are encoded into a RGB image.

A set of sparse fiducial points (landmarks) are selected and projected onto the image plane.

A deep convolutional network is trained to generate a heatmap for localizing the landmarks.





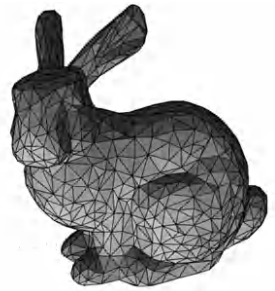
Gilani, Syed Zulqarnain, et al. "Deep, dense and accurate 3D face correspondence for generating population specific deformable models." Pattern Recognition 69 (2017): 238-250.

The problem of data

- Still, deep networks were rarely used in 3D vision until recently;
- This was mainly due to the fact that deep architectures need A LOT of data to be effectively trained;
- This because networks learn low-and-high level abstractions directly from the raw image data;
- → more data, better representation

The problem of data

- Until few years ago, there was a lack of 3D data available



- Nowadays, tons of 3D models are available in online repositories (e.g. 3D Warehouse)

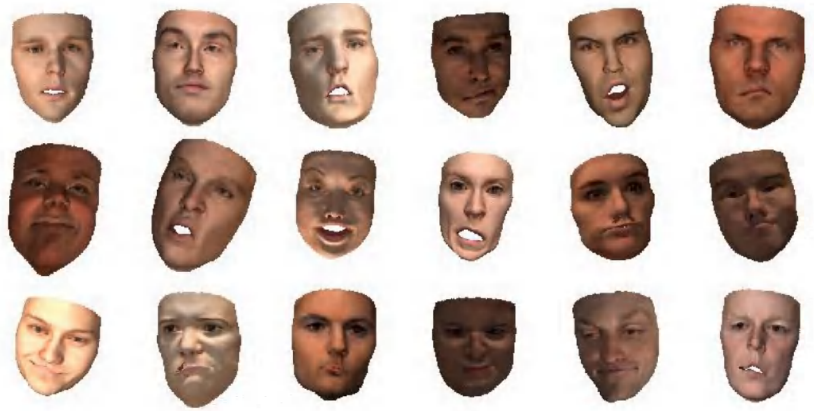
- For 3D faces, the problem still partially exists: collecting a large number of 3D face scans requires huge effort (and privacy related problems !)

- A widely used workaround consists in exploiting existing techniques to produce synthetic data to be used for training (like the 3DMM …)

3DMM For Generating Training Data

- One of the most popular applications of the 3DMM is the generation of synthetic training data.
- It can be used in very different tasks:
- In the work of Masi et al.[*], a 3DMM was used to render synthetic face images with diverse shapes and poses, and augment the training data of a face recognition deep network.
  - The 3DMM generated faces increase the invariance to slight shape changes

- Richardons et al. [**], instead use the 3DMM to render random RGB face images and train a reconstruction system.
  - The 3D geometry associated to each 2D image is known

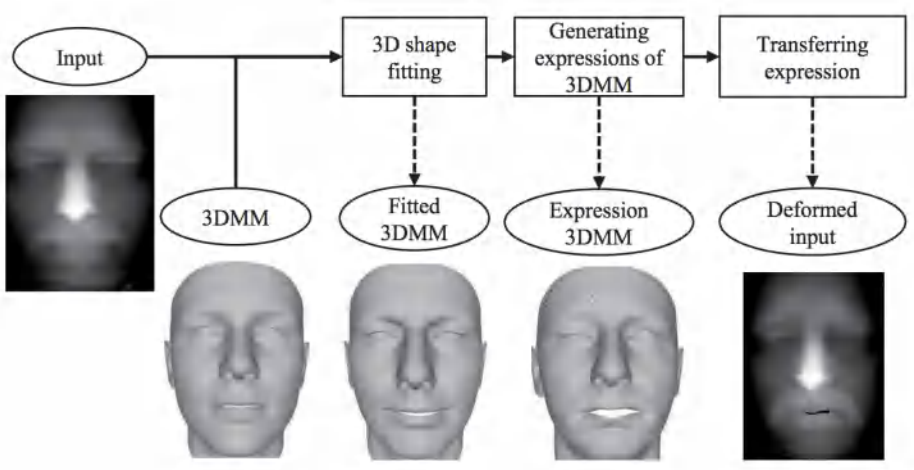


[*] Masi, Iacopo, et al. "Do we really need to collect millions of faces for effective face recognition?." European Conference on Computer Vision. Springer, Cham, 2016.

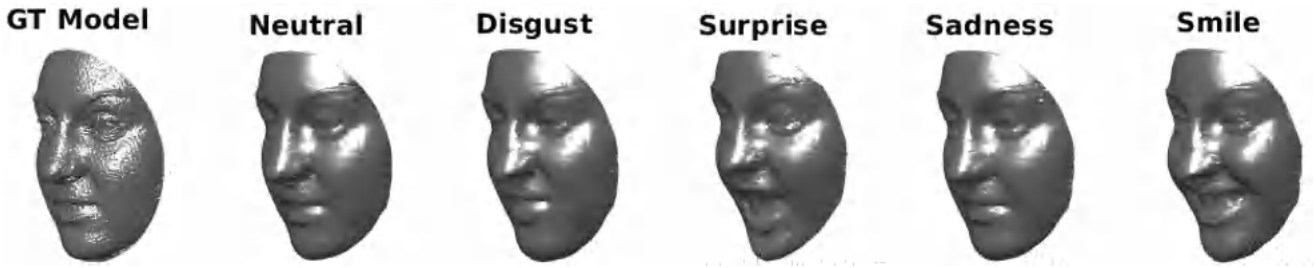
[**] Richardson, Elad, Matan Sela, and Ron Kimmel. "3D face reconstruction by learning from synthetic data." 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016.

3DMM For Generating Training Data

- The major limitation of using a 3DMM for generating data is that the possible plausible geometries that can be generated is constrained by the subspace spanned by the 3DMM parameters.
- However, the problem can be partially solved by using it in conjunction with real data.
- One can use both real and synthetic data or manipulate the real samples by, for example, transferring fake expressions.



Kim, Donghyun, et al. "Deep 3D face identification."
2017 IEEE international joint conference on biometrics (IJCB). IEEE, 2017.



Ferrari, Claudio, et al. "3DMM for Accurate Reconstruction of Depth Data.
" *International Conference on Image Analysis and Processing*. Springer, Cham, 2019.

A CNN for Regressing 3DMM Parameters

We have seen that the first 3DMM performed 3D face reconstruction by estimating a complex set of parameters to deform the 3DMM and render a synthetic image as similar as possible to the original.

So basically the problem was to find a mapping between pixels and the set of parameters…

Inspired by this, Tran et al.[*] used a CNN to regress this mapping

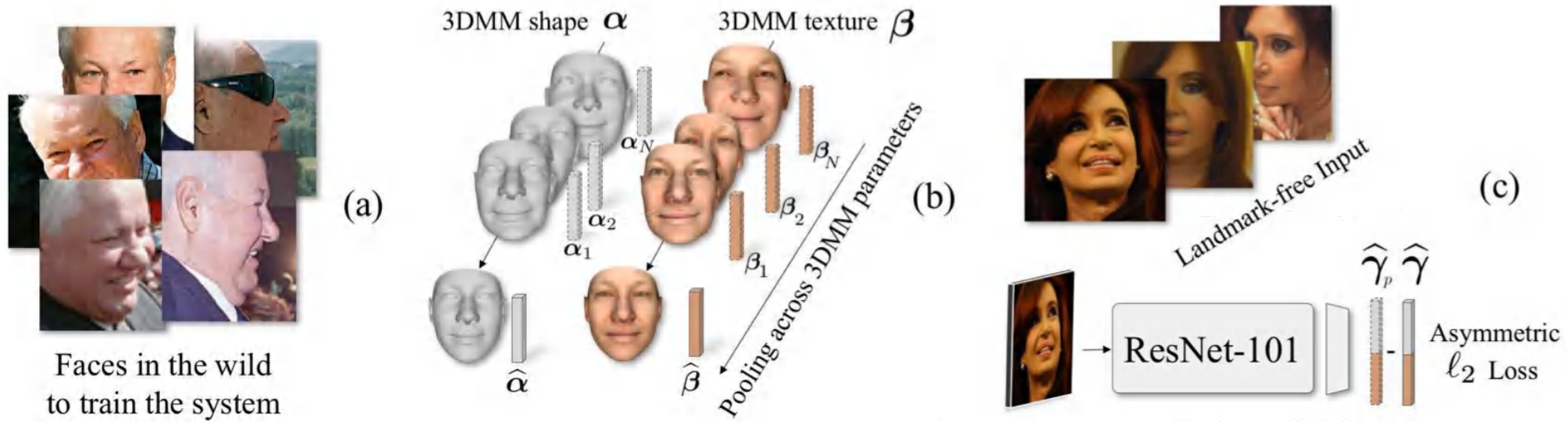
[*] Tran, Anh Tuan, et al. "Regressing robust and discriminative 3D morphable models with a very deep neural network." *CVPR, 2017*

A CNN for Regressing 3DMM Parameters

The 3DMM is applied to fit A LOT of 2D face images, clustered by identity.

For each identity, the 3DMM (shape and texture) parameters are pooled, assuming that the same subject should have the same set of parameters.

The parameters were used as signal to train the CNN and learn the mapping.

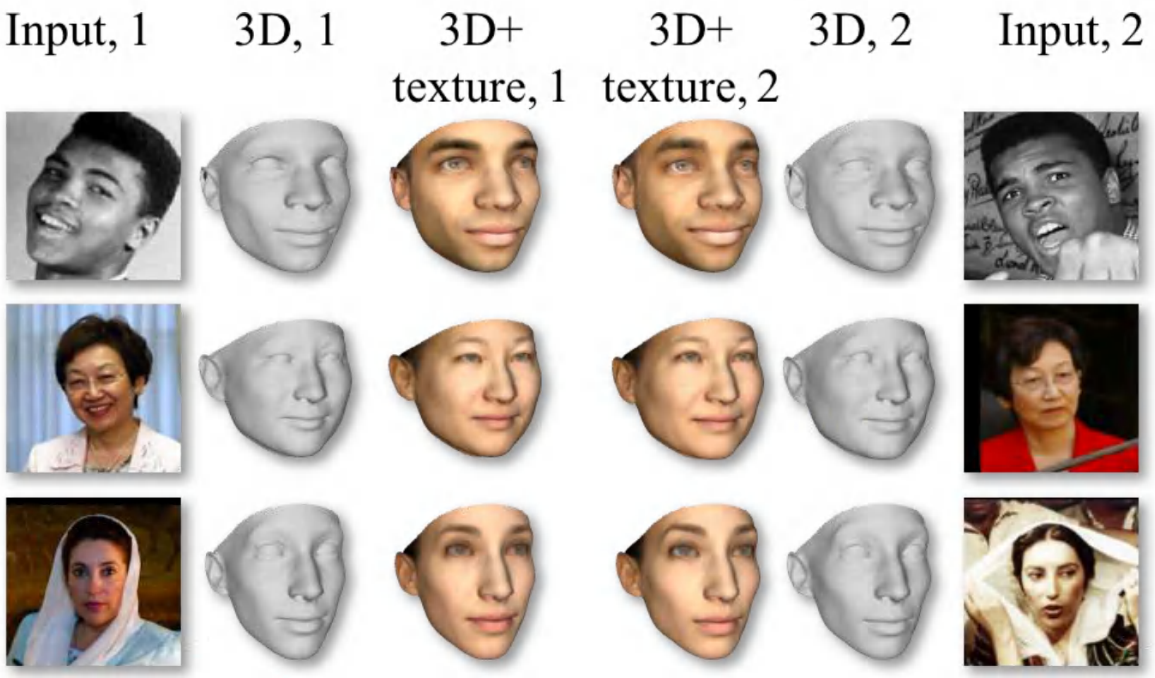


A CNN for Regressing 3DMM Parameters

The method demonstrated to be robust: images of the same subject led to very similar sets of parameters.

The estimated parameters have been also used to perform “in the wild” face recognition, with promising results

Again, the fidelity of the reconstruction is bounded by the modeling capabilities of the 3DMM!



→ Expressions are not reproduced!

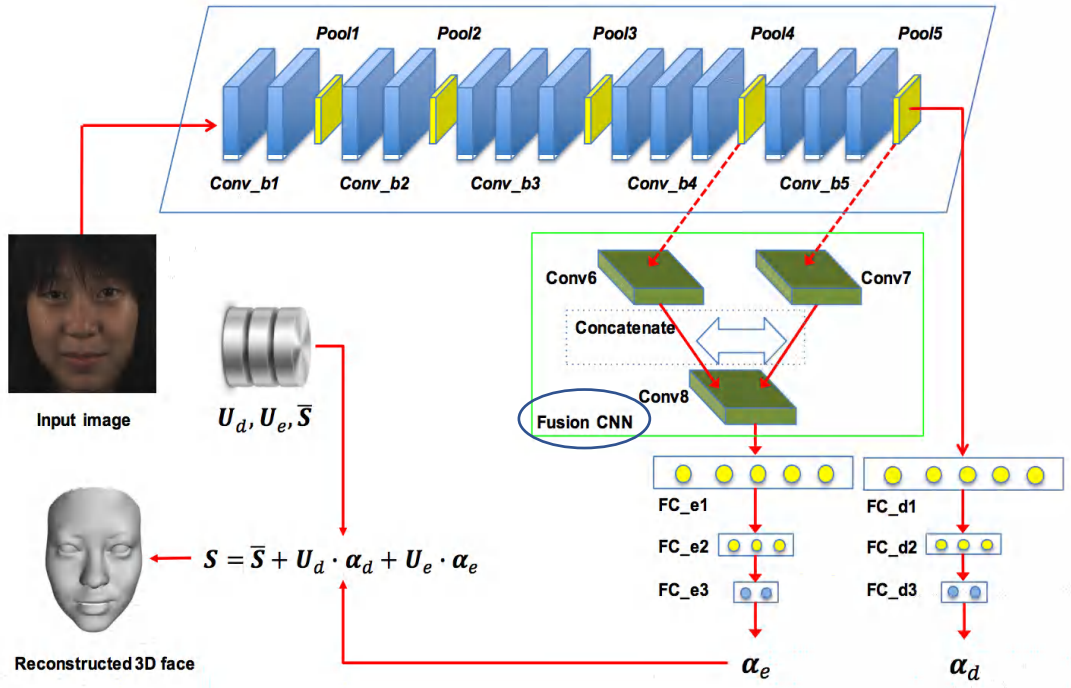
A CNN for Regressing 3DMM Parameters

A similar approach that focuses on 3D reconstruction and accounts for expressions is the one of Dou et al. [*]

They use a combination of two models, the BFM to model global shape deformations and the AFM [**] model to account for expressions

The CNN takes an RGB image as input and tries to regress the deformation parameters

The loss function is a composition of the identity and expression parameters



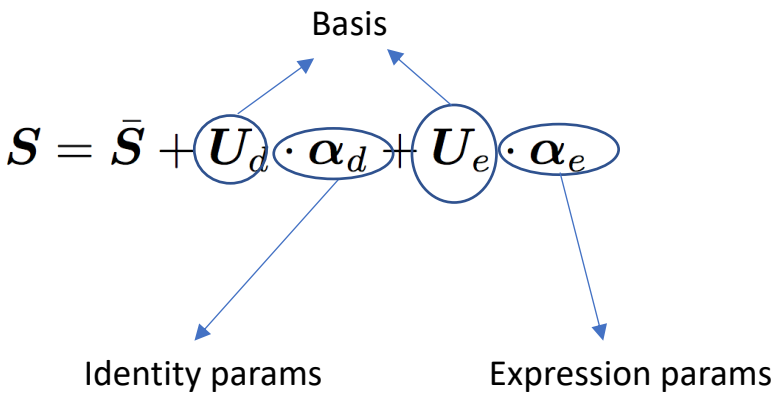
[*] Dou, Pengfei, et al. "End-to-end 3D face reconstruction with deep neural networks." *CVPR*. 2017.

[**] Kakadiaris, Ioannis A., et al. "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach." *IEEE TPAMI 2007*

A CNN for Regressing 3DMM Parameters

Also in this case, the 3DMM is used to generate synthetic training data to be used in conjunction with real data to train the network

The estimated parameters are used to generate the 3D shape associated to the image



Limitations

The methods seen so far reconstruct a “smooth” shape, without accounting for fine-grained details e.g. wrinkles.

Linear 3DMMs struggle in reproducing such fine details because of their intrinsic low dimensionality.

The power of CNNs can be exploited to recover highly detailed face surfaces

Also in this scenario, a shape prior (the 3DMM!) is useful to reconstruct the geometry of the face.

Tran, Anh Tuan, et al. "Extreme 3D Face Reconstruction: Seeing Through Occlusions." *CVPR*. 2018.

Fine-grained 3D Reconstruction

Many works employ the 3DMM to obtain a coarse, smooth reconstruction and then add fine details.

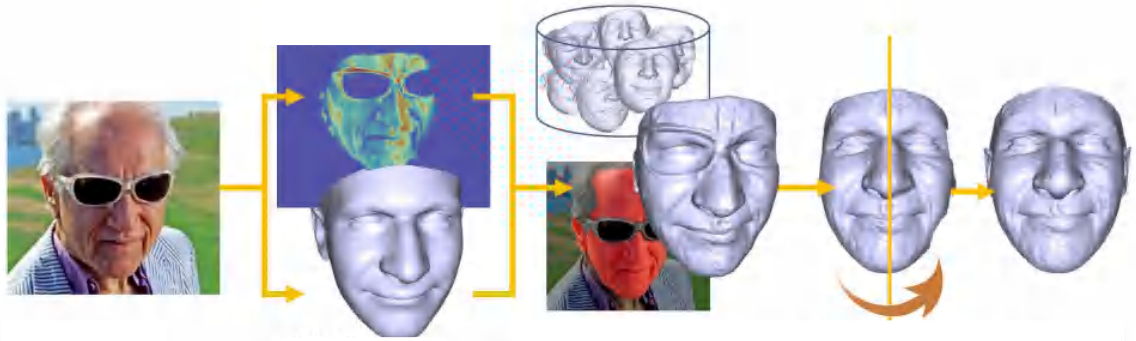
Tran et al. [*] refine a 3DMM-based reconstruction (foundation) using bump-maps *i.e.* depth differences.

The 3DMM is projected onto the image, such that it is aligned with the input face.

Depth displacements are added to make the foundation shape rendering match the input image

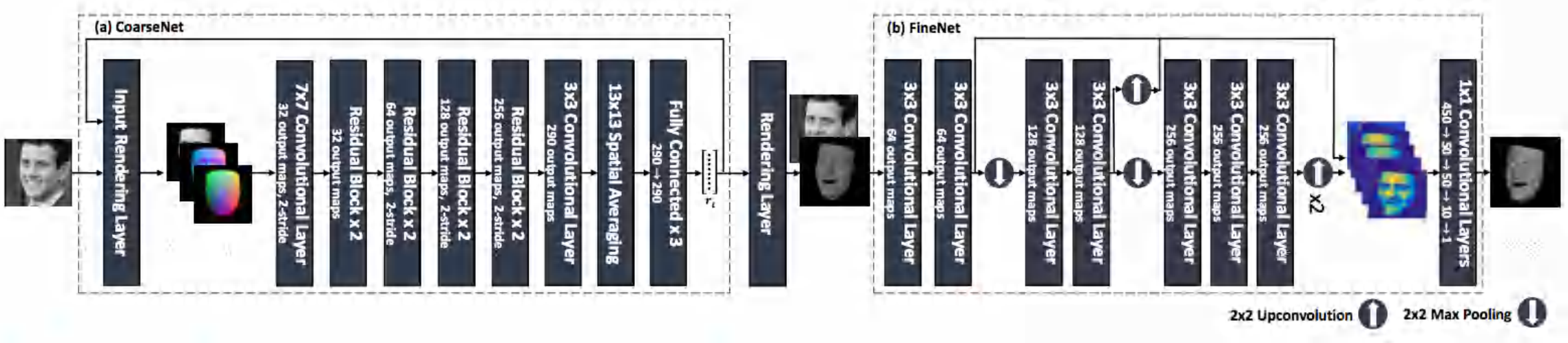
Using the 3DMM has the advantage of making the method robust to occlusions. In fact, the foundation shape is recovered deforming the 3DMM with the estimated parameters.

[*] Tran, Anh Tuan, et al. "Extreme 3D Face Reconstruction: Seeing Through Occlusions." *CVPR*. 2018.



Fine-grained 3D Reconstruction

Richardson et al. [5] developed a network to reconstruct a detailed face shape from single image in a coarse-to-fine manner



The first network (CoarseNet) is used to coarsely reconstruct the face shape

FineNet is then used to generate fine details

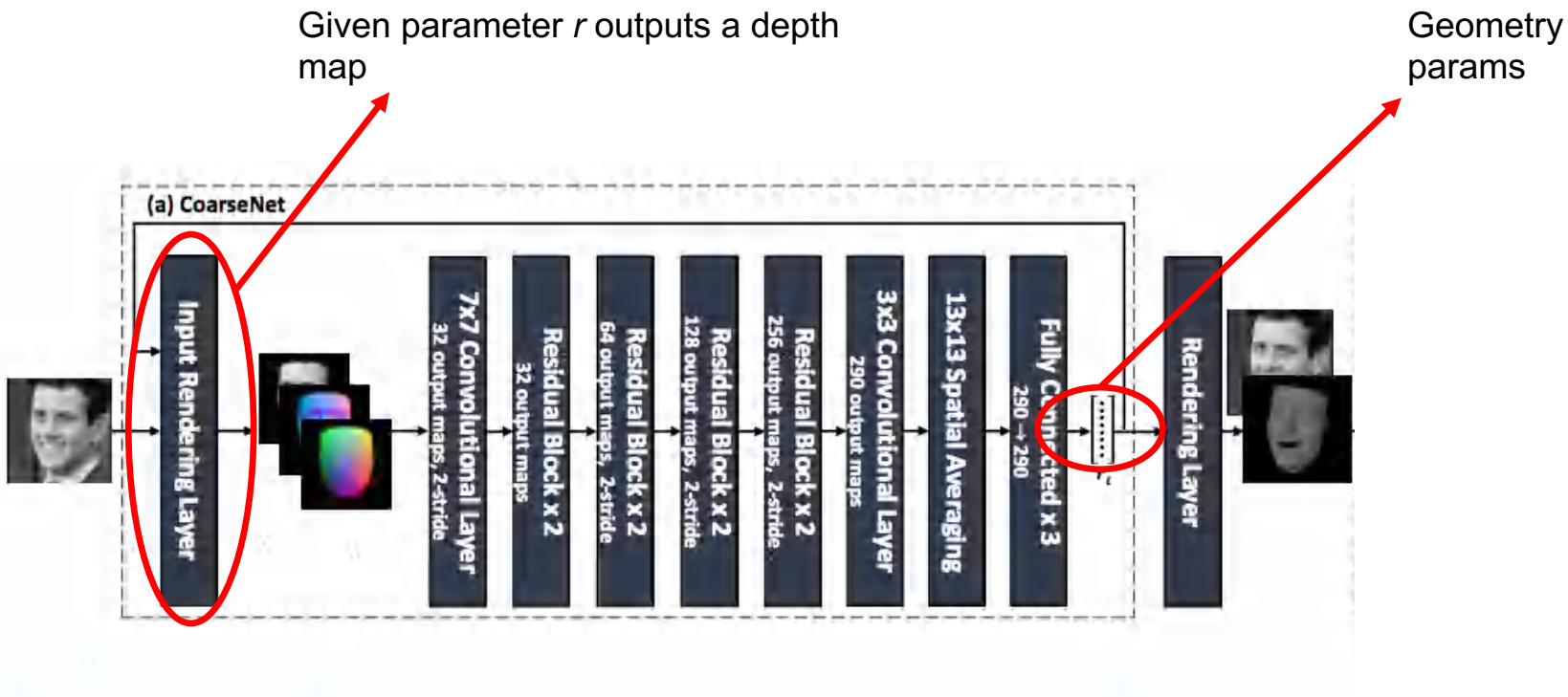
Also in this case, the training set for the CoarseNet is synthetically generated using the 3DMM; the training set consists of a **set of images** with associated **3DMM parameters**

[5] Richardson E., et al. " Learning Detailed Face Reconstruction from a Single Image " *CVPR, 2017*

Fine-grained 3D Reconstruction

Similarly to the previous works, CoarseNet is trained to regress the 3DMM and projection parameters (called *r*)

These parameters are used by a rendering layer to render the input image with the current set of parameters *r*, (image-parameters pair is the actual input of the network)

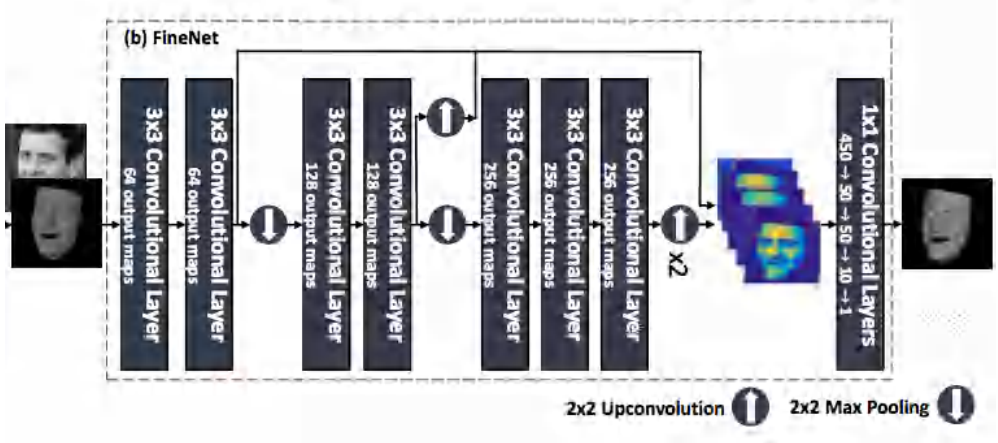


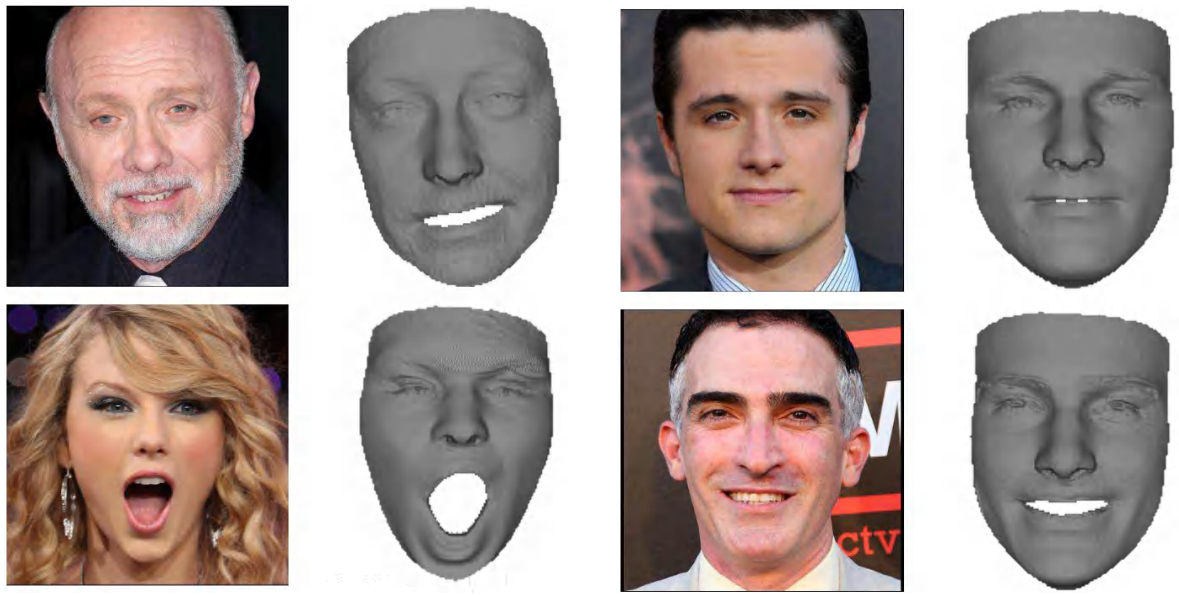
Fine-grained 3D Reconstruction

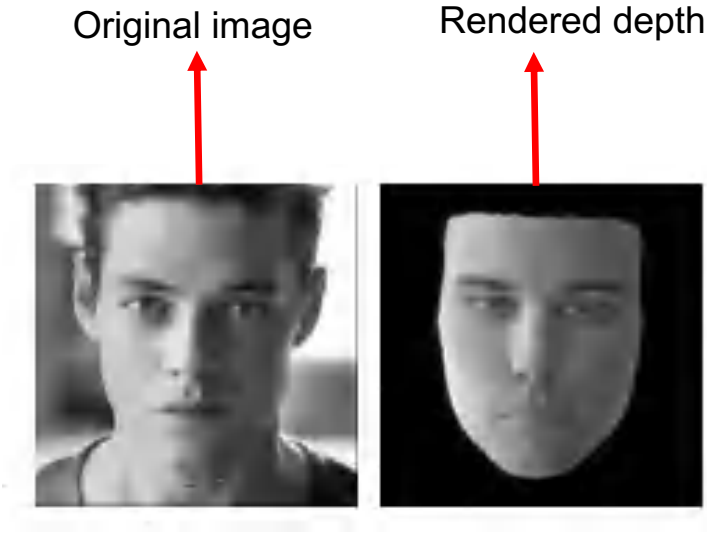
Finding sufficient ground-truth detailed surfaces to train FineNet is not possible → **unsupervised training**

The idea is to recover albedo and illumination parameters to render the recovered depth map

FineNet is then trained to refine the depth map in order to match the appearance of the input image

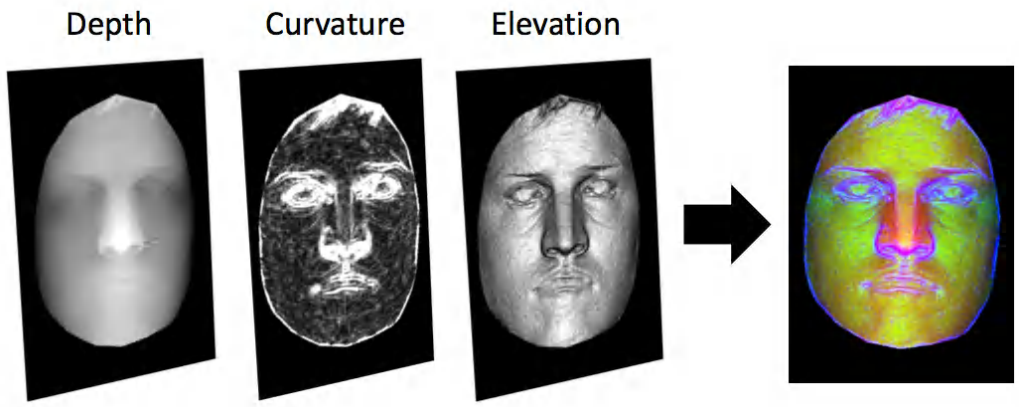


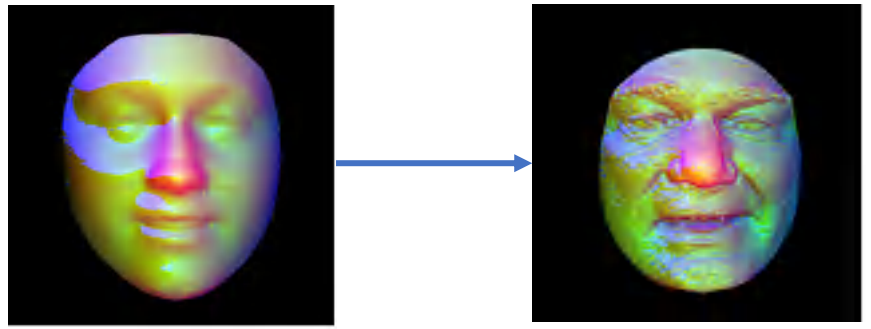




Our Contribution

- Recently, we developed a system for refining the coarse reconstruction provided by a 3DMM exploiting a Conditional Generative Adversarial Network (CGAN)
- In our system, the 3D models are rendered as RGB images where depth, elevation angle of the surface normals and curvature constitute the 3 channels.
- The curvature channel highlights significant face regions such as mouth, nose contour and eyes.
- The goal is to refine the coarse 3DMM reconstruction without exploiting the original RGB input image, making the solution independent from the input image.
- In this way, we ease the problem (but do not recover the texture !)

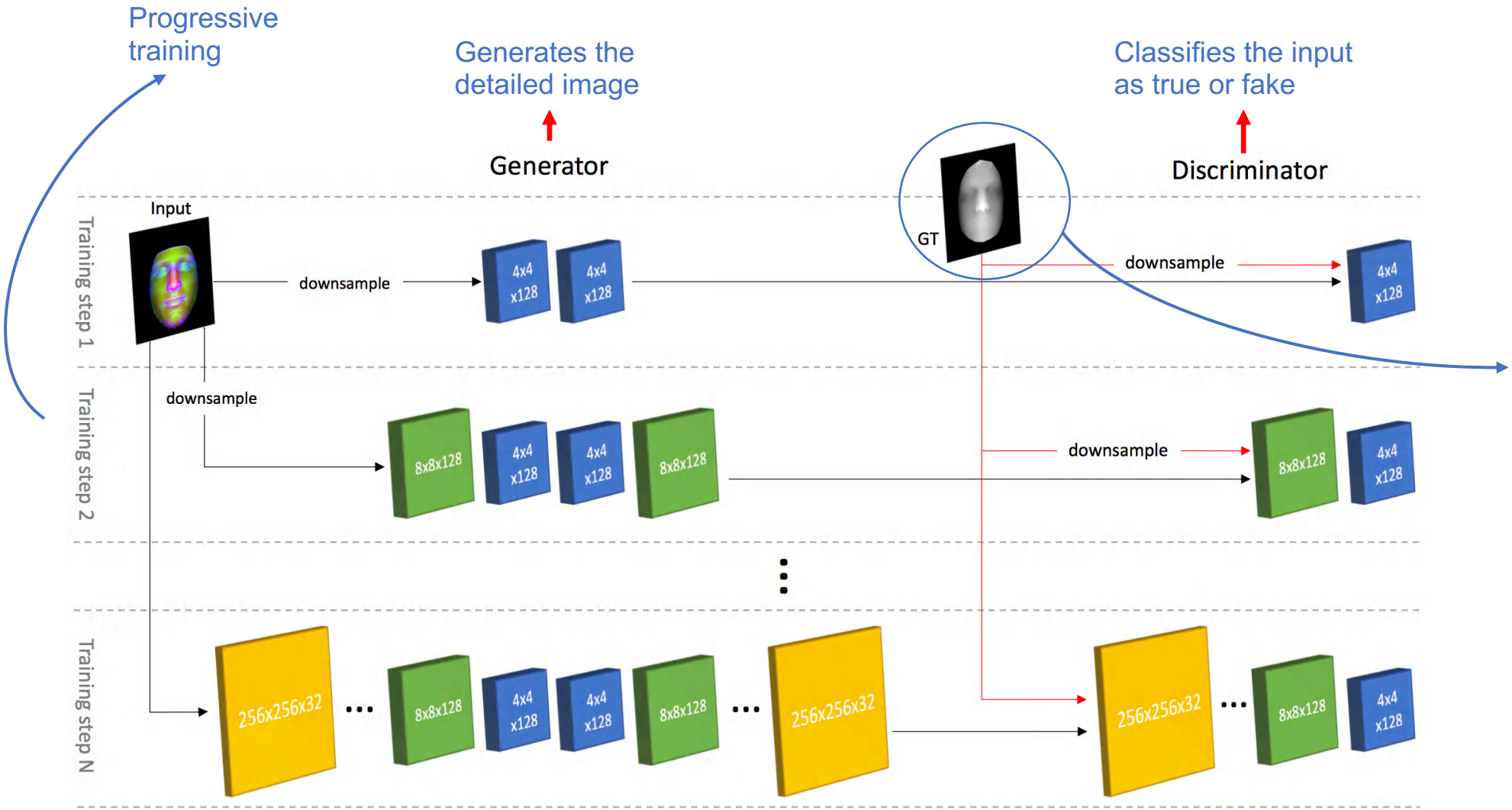




3DMM Refined

Galteri, Leonardo, et al. "Deep 3D morphable model refinement via progressive growing of conditional Generative Adversarial Networks." *Computer Vision and Image Understanding* 185 (2019): 31-42.

Deep 3DMM Refinement

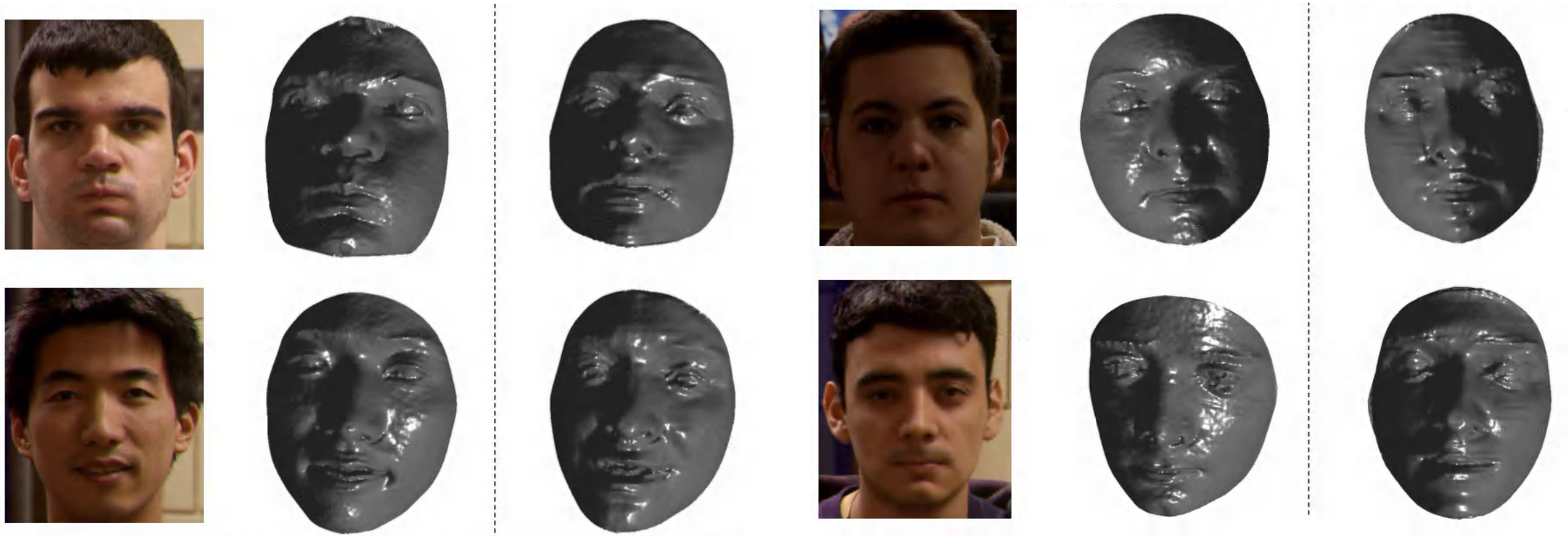


Only the depth channel is used to discriminate.

Curvature and normal vectors are subordinated to the 3D shape

Some Results

Input GT Reconstruction Input GT Reconstruction



Texture Matters

Other works exploit the non-linearity of deep networks to generate highly photorealistic textures.

The GANFIT method of Gecer et al. [*] exploit a GAN to generate highly realistic textures, that support the correct estimation of the 3D shape, obtained with a 3DMM that accounts for identity and expressions.

They also introduce the differentiable renderer, where camera and rendering parameters are learned as any other internal parameter of the network, allowing fast and accurate processing.

Landmarks consistency loss and identity preserving loss are also employed to guarantee consistent outputs

Gecer, Baris, et al. "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Not Only 3DMM Parameters

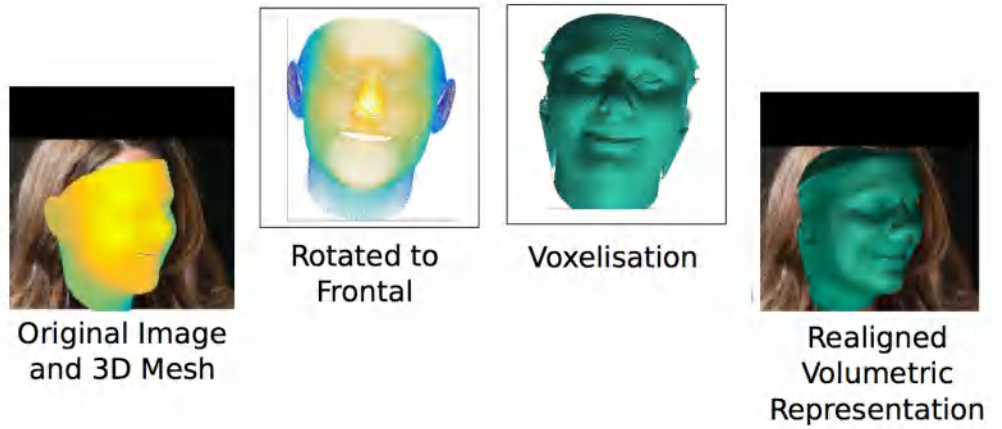
All the approaches seen so far either regress the 3DMM parameters from an image, or try to refine a 3DMM-based shape.

This is not the only option: an example is the work of Jackson et al.[*] in which a volumetric representation is regressed directly from the RGB image

Training 3D models (that are 3DMM-based reconstructions !!) are converted to a binary volume (voxelization) and the 3D space is discretized into a volume {H x W x D} = {192 X 192 X 200}

The goal is to learn a mapping from image to volume

$$f : \mathbf{I} \rightarrow \mathbf{V}$$

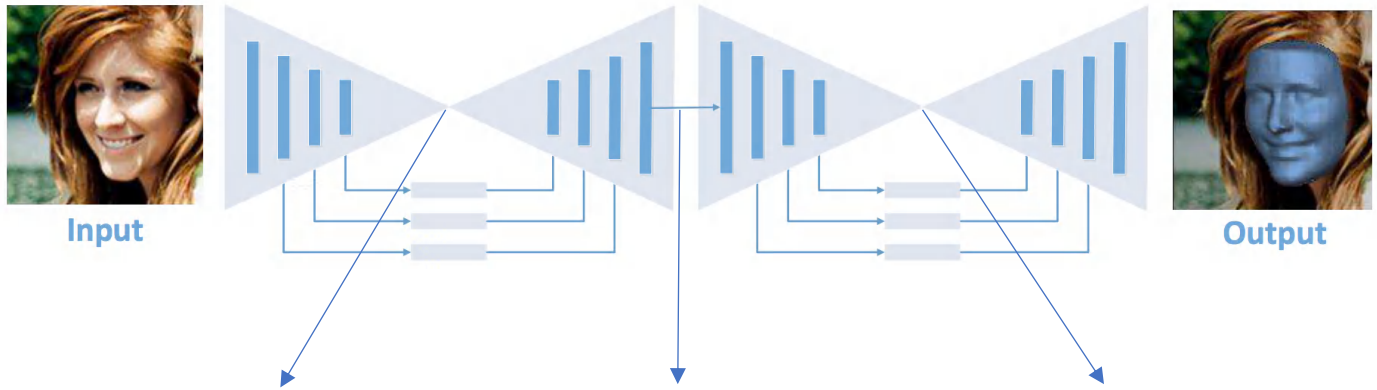


[4] Jackson, Aaron S., et al. "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression." *ICCV* 2017.

Volumetric Regression

The architecture is composed of two encoder-decoder networks that allow to maintain the spatial consistency between input and output. The 3DMM fitting step is bypassed.

Points enclosed by the 3D scan are given value 1, 0 otherwise.



The model is trained with a sigmoid cross entropy loss

Predicted value

GT value

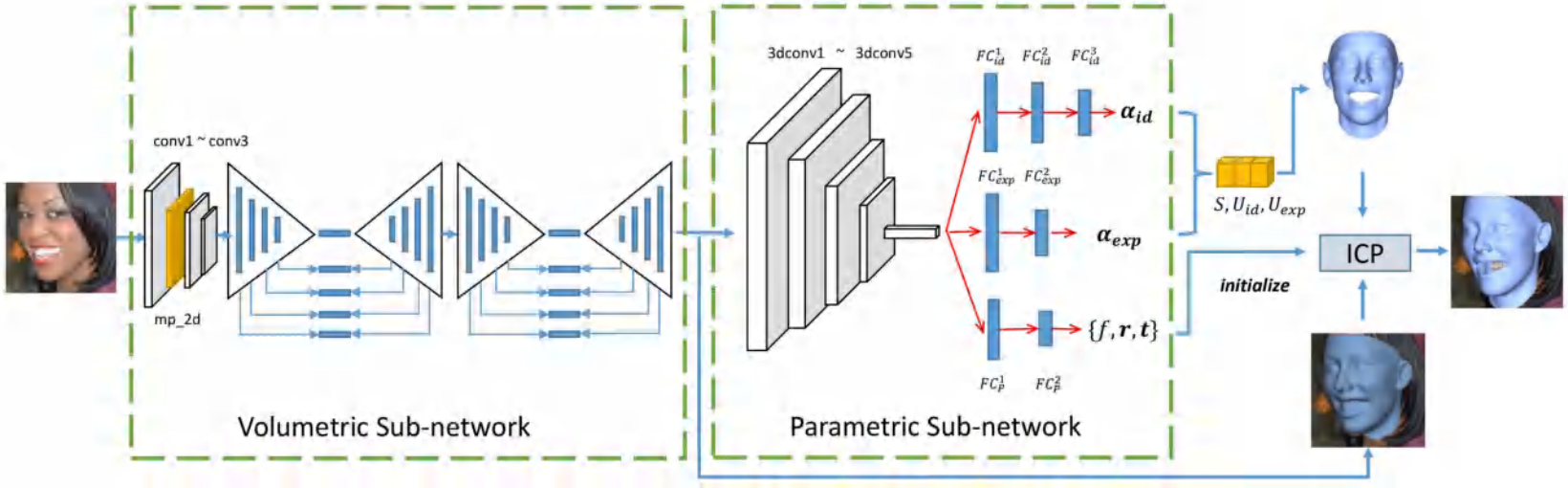
$$l_1 = \sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{d=1}^{D}[V_{whd}\log\hat{V}_{whd}+(1-V_{whd})\log(1-\hat{V}_{whd})]$$

Mixed Solutions

The latter solution allows bypassing the 3DMM fitting step. The resulting shapes are partially less constrained by the 3DMM shape space of possible geometries.

On the other hand, the reconstructed volumes have less resolution and cannot be further manipulated.

Yi et al. [*] proposed a mixed solution to address the issue, using both a volumetric sub-network to estimate a volume, and a parametric sub-network to transform the volume into a fully editable 3DMM.



Yi, Hongwei, et al. "Mmface: A multi-metric regression network for unconstrained face reconstruction." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

To summarize

Many other solutions exist and are constantly proposed.

Almost all of them rely on the 3DMM either to:

- Obtain sufficient training data
- Generalize to “in the wild” images, for which real 3D data does not exist
- Further manipulate the reconstruction
- And maybe others

We have seen that the 3DMM, despite being quite a “old” technique, is still very important and widely explored.

So far, we have seen how modern deep learning solutions can be exploited to improve the reconstruction

In the next part, we will see how deep learning can be used to learn more complex and descriptive non-linear 3D morphable models

UNIVERSITÀ
DEGLI STUDI
FIRENZE

Thanks!

Claudio Ferrari, Stefano Berretti, Alberto Del Bimbo

claudio.ferrari@unifi.it

https://sites.google.com/unifi.it/3dface-tutorial-cvpr20

Department of Information Engineering (DINFO) &
Media Integration and Communication Center (MICC)

University of Florence (UNIFI), Florence, Italy