Plug-and-Train Loss for Model-Based Single View 3D Reconstruction

Eduard Ramon
Crisalix SA
eduard.ramon@crisalix.com

Jordi Villar
Crisalix SA
jordi.villar@crisalix.com

Guillermo Ruiz
Crisalix SA
guillermo.ruiz@crisalix.com

Thomas Batard
Crisalix SA
thomas.batard@crisalix.com

Xavier Giró-i-Nieto
Universitat Politècnica de Catalunya
xavier.giro@upc.edu

Abstract

Obtaining 3D geometry from images is a well studied problem by the computer vision community. In the concrete case of a single image, a considerable amount of prior knowledge is often required to obtain plausible reconstructions. Recently, deep neural networks in combination with 3D morphable models (3DMM) have been used in order to address the lack of scene information, leading to more accurate results. Nevertheless, the losses employed during the training process are usually a linear combination of terms where the coefficients, also called hyperparameters, must be carefully tuned for each dataset to obtain satisfactory results. In this work we propose a hyperparameters-free loss that exploits the geometry of the problem for learning 3D reconstruction from a single image. The proposed formulation is not dataset dependent, is robust against very large camera poses and jointly optimizes the shape of the object and the camera pose.

1. Introduction

3D technology is key for a wide range of industries. Medicine, construction, cinema and many other disciplines can nowadays digitalize the world we perceive using 3D reconstruction algorithms, create new objects by means of 3D printers or analyze the world using 3D segmentation techniques. The methods used for reconstructing 3D objects range from highly accurate scanners based on photometry, which are voluminous and expensive, to scanners based on structured light, which may be less precise but portable and much cheaper. These solutions, yet valid, do not reach the mainstream users, since they require specific hardware. Designing new algorithms capable of reconstructing 3D objects precisely from a single or multiple RGB images will help democratizing 3D technology, allowing more people to take advantage of its possibilities.

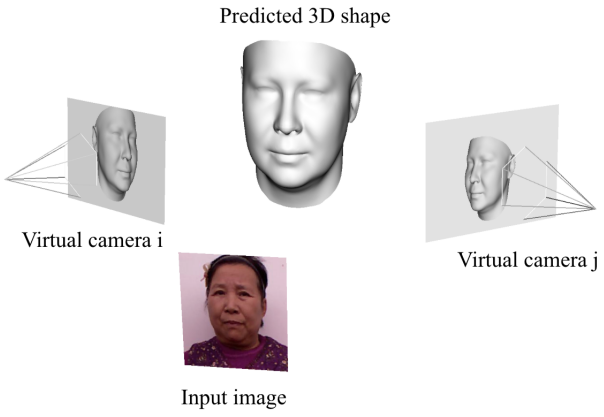

Figure 1: Overview of our random projections approach for implicit 3D shape regularization.

Inferring the geometry of objects from a single or multiple images is a well-studied problem by the computer vision community. Traditionally, the employed techniques have been based in geometry and/or photometry [12, 32], which usually require a large amount of images in order to create precise reconstructions. Recently, the capacity of deep neural networks [9] to obtain hierarchical representations of the images and to encode prior knowledge has been applied to 3D reconstruction in order to learn the implicit mapping between images and geometry [7, 33].

Nevertheless, employing deep neural networks to solve 3D related problems implies some issues that need to be addressed. One of the main drawbacks is the 3D data representation. The trivial generalization from 2D images to 3D space are the 3D voxel grids. This representation, which is simple and allows the use of 3D convolutions, does an inefficient use of the target space when trying to reconstruct surfaces. Moreover, state of the art methods that use this representation mostly work at resolutions around 128x128x128

voxels [7, 33], which are too small for most of the applications. 3D meshes [15, 31] are a more convenient representation because they efficiently model surfaces and can be easily textured and animated for computer graphics applications. However, 3D meshes are defined in a non-Euclidean space, where the usual deep learning operations like convolutions are not defined. Geometric deep learning [3] is nowadays a hot research area to bring basic operations to non-Euclidean domains like graphs and manifolds, which is the case of 3D meshes. Finally, 3D Morphable Models (3DMM) [2] are used for category-specific problems to reduce the dimensionality of plausible solutions and lead to more robust and likely predictions.

Another challenge when working on 3D reconstruction using deep learning is the lack of labelled data. In tasks like image recognition, there exist large annotated datasets with millions of images [8]. Unfortunately, the data is not as abundant in 3D as it is in 2D and, consequently, researchers have walked around this limitation with different strategies. Defining losses in the image domain [29, 23] is a common approach since it provides flexibility to use different kinds of annotations like sparse sets of keypoints, foreground masks or pixel intensities. A second strategy is the use of synthetic data [22, 23, 25] since it provides perfect 3D groundtruth. Unfortunately, the systems trained with synthetic data tend to suffer from poor generalization due to the distribution gap between the training and the testing distributions.

Finally, subject to the 3D data representation and the availability of labels, several works have proposed different losses to learn their models [7, 33, 15, 31, 29, 23]. These losses usually present a number of terms related by weighting hyperparameters that need to be tuned for an effective optimization. Nevertheless, finetuning these parameters for each reconstruction dataset is a hard and computationally expensive task that presents high chances of achieving suboptimal results.

The main contributions of our work are:

- The Multiview Reprojection Loss (MRL), a novel single term loss for learning model-based 3D reconstruction from one image that does not require tuning any hyperparameter.
- A notable decrease of time and complexity for training single view 3D reconstruction models using real or synthetic data.
- A qualitative evaluation on FaceWarehouse dataset [4] and a quantitative one on MICC dataset [1] where we obtain comparable results to the state of the art.

MRL can be plugged in any model-based 3D reconstruction system. In this work, focus on the problem of learning single view 3D face reconstruction using 3DMM from real or synthetic 3D data. We show how MRL outperforms the standard multiterm losses, obtain excellent qualitative results on the neutral expressions of the Facewarehouse dataset [4], and achieves state of the art results in the MICC Florence 3D Faces dataset [1].

The rest of the paper is structured as follows. Section 2 reviews the state of the art for 3D reconstruction from a single image using deep learning models. Section 3 presents the novel single term Multiview Reprojection Loss and describes it geometrically. Section 4 compares MRL performance with respect to popular 3D reconstruction methods. Finally, Section 5 draws the conclusions of our work. Supplementary material can be found online at https://imatge-upc.github.io/mrl/.

2. State of the art

Classical methods used to obtain 3D reconstructions from multiple views are based either on multiview geometry or on photometric stereo. Geometry based techniques aim to estimate the 3D position of surface points from the object. Usually, the first stage of a geometric reconstruction process is a keypoint detector. Then, the detected keypoints are matched using handcrafted image descriptors like SIFT [21] or ORB [26]. Finally, multiview geometry is employed to formulate the reprojection error, which can be minimized using non-linear least squares algorithms like Levenberg-Marquardt. These techniques, also known as Structure from Motion (SfM), are appropriate for reconstructing rigid objects with detailed textures and when several views are available. On the other hand, photometric based techniques focus on estimating the normals of the surface of the objects by observing them under different light conditions. This group of algorithms, also known as Photometric Stereo or Shape from Shading (SfS) [12], make strong assumptions about the physical properties of the surface of the target objects, which often are not fulfilled.

Since AlexNet [20] succeeded in training a convolutional neural network (CNN) for large scale image recognition, multiple computer vision tasks have been tackled with deep neural networks [9]. Among them, 3D reconstruction has also benefited from their learned representations, obtaining important performance gains with respect to hand-crafted classic techniques. In general, two big groups of learning-based 3D reconstruction methods can be differentiated by the fact of using or not a 3DMM, which we define as model-based and model-free approaches respectively.

2.1. Model-free approaches

Methods that do not include a 3DMM in their core [33, 10, 31, 13, 16, 15], also called model-free, are usually oriented to solve more generic problems, such as reconstructing objects with different shapes, and are highly conditioned by the 3D representation they use.

For instance, methods based on 3D voxel grids [33, 10, 13] tend to use binary cross entropy as objective to optimize their architecture. Eventually, 3D voxel grid geometries can be projected into the image plane to construct supervision signals defined in the image domain, such as depth errors [16] or binary masks errors [33]. Despite their flexibility, 3D voxel grid methods are very inefficient at representing surfaces, and hierarchical models are required to achieve denser representations [10]. Although they have been mostly assessed in synthetic datasets [6], 3D voxel grid methods have also obtained state of the art results in real applications [13].

Meshes are a common alternative to 3D voxel grids since they are more efficient at surface modelling and have more potential applications. Recent works [15, 31] suggest that state of the art results can be achieved by minimizing the Chamfer Loss while regularizing the surface through the Laplace-Beltrami operator and other geometric elements such as normals [31]. In addition, a family of novel and relevant operators that have been successfully applied to 3D reconstruction with meshes [31] are the Graph Convolutional Networks (GCN) [3], which generalize the convolution operator to non-Euclidean domains.

2.2. Model-based approaches

Model-free methods, specially the mesh based approaches, need to be heavily regularized by using geometric operators in order to obtain plausible 3D reconstructions and, despite its flexibility, they are difficult to train. Model-based approaches offer a simpler solution to regularize surfaces by modeling them as a linear combination of a set of basis [2]. Thus, the learning problem is reduced to estimate a vector of weights to linearly combine the basis of the model.

Due to the lack of 3D data, some works have driven their experiments towards the evaluation of models trained on synthetic data [22] [23]. Yet obtaining successful results, iterative error feedback (IEF) [5] is usually required for good generalization, which unfortunately implies multiple passes through the network. To speed up the IEF, [14] performs this process in the latent space. Since using synthetic data provides perfect labels, the losses are designed to explicitly model the error between predictions and groundtruth model parameters. Moreover, regularization is added as another loss term by enforcing the norm of the 3DMM parameters to be small.

On the other hand, some methods overcome the scarcity of 3D data by defining losses directly in the image domain [29, 28, 23]. This avoids using IEF since the data is trained and tested in the same distributions. On the other hand, annotations on the image domain are required [34] or differentiable renderers [17] are necessary to construct self-supervised losses using the raw pixel values [29]. Again, regularization is needed on the predicted model weights to ensure the likelihood of the predicted 3D shapes.

A common feature of most methods used for learning 3D reconstruction is the need of some sort of regularization applied to the predicted 3D shape. Regularization is added as a weighted combination of terms in the loss, either geometric operators for meshes, or norms of the predicted shape model parameters for model-based approaches. These terms provide the model with stability but, at the same time, add complexity to the model and consequently to its optimization. In [14], an adversarial regularization is proposed in order to penalize predicted samples that fall out of the target distribution. This statistical approach is more generic and simple.

Our work follows the direction of [14] with the objective of finding more generic losses to learn model-based single view 3D reconstruction that simplify the optimization of the architectures. In contrast to them, we propose a geometric approach, instead of a statistical one, that fuses the data terms and the regularization terms into a single objective held by the geometry of the problem. In this way, we eliminate all the hyperparameters of the loss.

3. Mutiview Reprojection Loss

In this section we present our MRL, a loss without hyperparameters for learning single view 3D reconstruction. Firstly, we describe the elements of a single view learning-based reconstruction problem. Then, a detailed explanation is given on how data and regularization terms can be fused into a single term using geometry in order to learn 3D shape and pose simultaneously.

3.1. Problem statement

A learning-based single view 3D reconstruction problem can be defined as finding the unknown mappings from an input image $\mathcal{I}$ to a 3D shape $\boldsymbol{x} \in \mathbb{R}^{3N}$, being N the number of points, and to the camera pose $c = [R|\boldsymbol{t}]$, R being the rotation of the camera and $\boldsymbol{t} = (t_x, t_y, t_z) \in \mathbb{R}^3$ the spatial position of the camera. We model the rotation of the camera R as a unit quaternion $\boldsymbol{q} = (q_0, q_1, q_2, q_3) \in \mathbb{H}_1$ to avoid the Gimbal lock effect, which is the loss of one degree of freedom in a three-dimensional mechanism.

We split the mappings to be learned in four functions: $\mathcal{E}$, $\mathcal{X}$, $\mathcal{Q}$ and $\mathcal{T}$. The former function $\mathcal{E}$ is intended to extract relevant features from $\mathcal{I}$ and the rest to map these features to $\boldsymbol{x}$, $\boldsymbol{q}$ and $\boldsymbol{t}$ respectively, so that $\hat{\boldsymbol{x}} = \mathcal{X}(\mathcal{E}(\mathcal{I}))$, $\hat{\boldsymbol{q}} = \mathcal{Q}(\mathcal{E}(\mathcal{I}))$ and $\hat{\boldsymbol{t}} = \mathcal{T}(\mathcal{E}(\mathcal{I}))$.

Most of the current methods based on deep neural networks [22, 23] learn the mapping functions $\mathcal{E}$, $\mathcal{X}$, $\mathcal{Q}$ and $\mathcal{T}$ by linearly combining different loss terms.

Each of these terms is responsible for controlling a property of the reconstruction, and its contribution to the final loss is adjusted by a weighting hyperparameter that must be

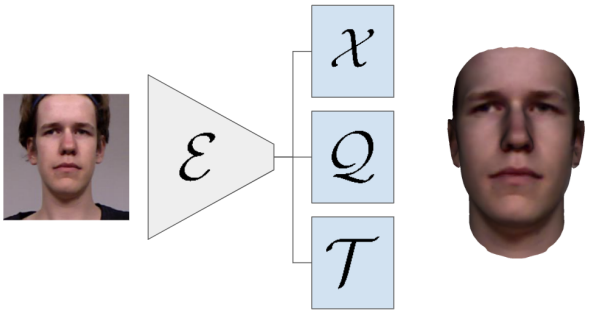

Figure 2: Modules used in the architecture to evaluate the proposed single term loss.

tuned. In general, these loss terms can be divided in data terms and regularization terms [29].

Data terms are the ones that guide the network predictions towards matching the ground truth labels during training. In the considered 3D reconstruction setup, these correspond to the the 3D shape $\boldsymbol{x}$ and the camera pose $\{\boldsymbol{q}, \boldsymbol{t}\}$:

$$\mathcal{L}_{data} = \mathcal{L}_x + \alpha\mathcal{L}_q + \beta\mathcal{L}_t. \tag{1}$$

As noted in [18], the relation between hyperparameters α and β varies substantially depending on the problem and, consequently, the choice of these parameters has a severe impact for the camera pose estimation.

On the other hand, regularization terms control the resulting 3D shape $\boldsymbol{x}$ in terms of geometric and semantic likelihoods. In this sense, using a 3DMM allows to control geometry and semantics in a lower dimensional space. More precisely, the use of a 3DMM enables the definition of a 3D shape as:

$$\boldsymbol{x} = \boldsymbol{m} + \Phi_{id}\boldsymbol{\alpha}_{id}, \tag{2}$$

where $\boldsymbol{m}$ represents the mean of $\boldsymbol{x}$, and Φ_{id} and $\boldsymbol{\alpha}_{id}$ are the identity basis and the identity parameters respectively.

In order to obtain plausible shapes, $\boldsymbol{\alpha}_{id}$ needs to have a small norm. Consequently, losses include an extra regularization term that force this condition during training:

$$\mathcal{L}_{reg} = \gamma||\boldsymbol{\alpha}_{id}||_2^2. \tag{3}$$

The final loss simply sums the data and regularization terms:

$$\mathcal{L} = \mathcal{L}_{data} + \mathcal{L}_{reg}. \tag{4}$$

Following this multiterm driven strategy, most of the methods for 3D reconstruction need to estimate the weighting hyperparameters for each specific dataset, a hard and expensive process that might lead to suboptimal results.

3.2. A unique term loss for single view 3D reconstruction

We introduce a novel formulation of the loss for learning 3D reconstruction from a single image that does not require any weighting hyperparameter. We get inspiration from [18], where they propose to unify the data term into a single expression that exploits the geometry of the scene for the task of camera pose estimation. The proposed single view reprojection loss is defined as:

$$\mathcal{L}_{data} = ||\mathcal{P}(\boldsymbol{q}, \boldsymbol{t})(\boldsymbol{x}) - \mathcal{P}(\hat{\boldsymbol{q}}, \hat{\boldsymbol{t}})(\hat{\boldsymbol{x}})||_1, \tag{5}$$

where $\mathcal{P}$ represents a projective transform from 3D to the 2D image plane:

$$\begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} = K[R(\boldsymbol{q})|\boldsymbol{t}]\boldsymbol{x}_H \tag{6}$$

$$\boldsymbol{x}_{2D} = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u'/w' \\ v'/w' \end{pmatrix}, \tag{7}$$

being K the calibration matrix, $R(\boldsymbol{q})$ the rotation matrix induced by the quaternion $\boldsymbol{q}$ and $\boldsymbol{x}_H$ the shape in homogeneous coordinates.

This formulation, introduced in [18], elegantly unifies learning position and rotation within a single term. In contrast to [18], our loss aims to predict the 3D shape $\boldsymbol{x}$ as well. In order to reduce the number of parameters, we exploit the 3DMM-oriented shape from Equation 2 to develop Equation 5 as:

$$\begin{aligned}\mathcal{L}_{data} = ||&\mathcal{P}(\boldsymbol{q}, \boldsymbol{t})(\boldsymbol{m} + \Phi_{id}\boldsymbol{\alpha}_{id}) - \\ &\mathcal{P}(\hat{\boldsymbol{q}}, \hat{\boldsymbol{t}})(\boldsymbol{m} + \Phi_{id}\hat{\boldsymbol{\alpha}}_{id})||_1.\end{aligned} \tag{8}$$

By using Equations 5 or 8 as losses, one can simultaneously learn shape and pose by minimizing the reprojection error. Unfortunately, optimizing 3D shape and pose by projecting into a single image plane is not possible without regularization. As it can be observed in Figure 3, the network learns to generate flattened shapes $\hat{\boldsymbol{x}}$, which produce minimum reprojection error, but at the same time are not likely to belong to the distribution of 3D facial shapes.

3.3. Implicit regularization via random projections

In a multiterm loss like Equation 1, a trivial solution to regularize the predictions of $\hat{\boldsymbol{x}}$ would be to add an extra term, $||\hat{\boldsymbol{\alpha}}_{id}||_2^2$, to keep the norm of $\hat{\boldsymbol{\alpha}}_{id}$ small. This would introduce an extra hyperparameter that we would like to avoid.

Instead, we propose to implicitly regularize the learning process of $\hat{\boldsymbol{x}}$ by projecting to multiple image planes. To do so, we modify the projection function $\mathcal{P}$ to include a

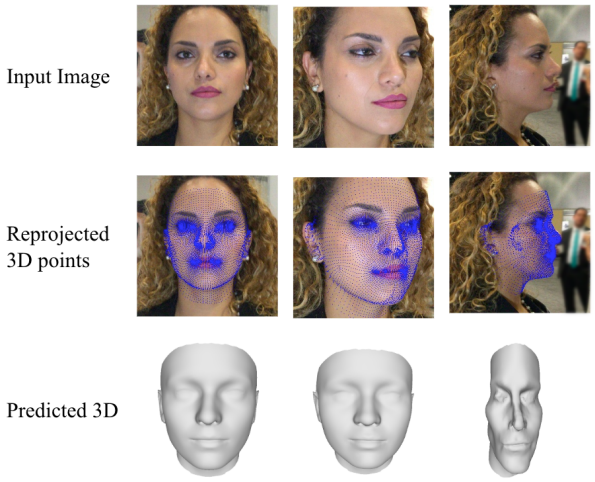

Figure 3: Effect of training with Equation 5. While the reprojection error is minimized, the 3D shape is not plausible.

distortion matrix D that reformulates the projection matrix as:

$$\mathcal{P}_D(\boldsymbol{q}, \boldsymbol{t}) = K[R(\boldsymbol{q})|\boldsymbol{t}]D\boldsymbol{x}, \tag{9}$$

and rewrite the single term loss from Equation 8 to project to multiple random views:

$$\mathcal{L} = \sum_{v=1}^{V} ||\mathcal{P}_I(\boldsymbol{q_v}, \boldsymbol{t_v})(\boldsymbol{x}) - \mathcal{P}_D(\boldsymbol{q_v}, \boldsymbol{t_v})(\hat{\boldsymbol{x}})||_1, \tag{10}$$

where $\boldsymbol{q_v}$ and $\boldsymbol{t_v}$ represent the camera pose of a random view, I is the identity matrix and D is a distortion matrix caused by the difference between $\boldsymbol{q}, \boldsymbol{t}$ and $\hat{\boldsymbol{q}}, \hat{\boldsymbol{t}}$, which is computed as $D = [R(\boldsymbol{q})|\boldsymbol{t}] \cdot [R(\hat{\boldsymbol{q}})|\hat{\boldsymbol{t}}]^{-1}$.

In practice, we find that the optimization process is more stable when the camera pose parameters $\{\hat{\boldsymbol{q}}, \hat{\boldsymbol{t}}\}$ and the 3D shape $\hat{\boldsymbol{x}}$ are related by addition instead of product. Thus, we reformulate Equation 10 as:

$$\begin{aligned}\mathcal{L} = \sum_{v=1}^{V} &||\mathcal{P}_I(\boldsymbol{q_v}, \boldsymbol{t_v})(\boldsymbol{x}) - \mathcal{P}_I(\boldsymbol{q_v}, \boldsymbol{t_v})(\hat{\boldsymbol{x}})||_1 \\ &+||\mathcal{P}_I(\boldsymbol{q_v}, \boldsymbol{t_v})(\boldsymbol{x}) - \mathcal{P}_D(\boldsymbol{q_v}, \boldsymbol{t_v})(\boldsymbol{x})||_1.\end{aligned} \tag{11}$$

It can be easily proved that, using Equation 11, the gradients of loss with respect $\hat{\boldsymbol{x}}$, $\nabla_{\hat{\boldsymbol{x}}}\mathcal{L}$, do not depend on the error caused by $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{t}}$, and vice versa. We think that this is the reason why training with Equation 11 produces a more stable optimization than using Equation 10.

We define the expression in Equation 11 as the *Multiview Reprojection Loss (MRL)*, which represents the main contribution of our work. MRL allows to simultaneously learn the 3D shape and the camera pose without explicit regularization of $\hat{\boldsymbol{x}}$, since it penalizes the error signals consistently.

4. Experiments

This section evaluates MRL in in terms of accuracy and robustness. Section 4.1 describes the architecture of the convolutional neural network trained with MRL in a private dataset. The trained model is evaluated from three different perspectives: Section 4.2 compares in terms of reprojection loss MRL with a multiterm loss, Section 4.3 show the robustness of the model to diverse face appearances, while the results on the public FaceWareHouse [4] and MICC [1] datasets are reported in Sections 4.4 and 4.5, respectively.

4.1. Implementation details

We make use of a standard architecture to predict the first 50 identity parameters $\boldsymbol{\alpha}_{id}$ of a 3DMM, the camera rotation as a quaternion $\boldsymbol{q} = (q_0, q_1, q_2, q_3)$ and the spatial camera translation $\boldsymbol{t} = (t_x, t_y, t_z)$. Similarly to [22, 29, 28, 23] we choose a convolutional neural network as encoder $\mathcal{E}$ based on VGG-16 [27] to extract image features, and then three multiyer perceptrons (MLP), $\mathcal{S}$, $\mathcal{Q}$ and $\mathcal{T}$, with 1 hidden layer of 256 units, that are added on top of $\mathcal{E}$ to regress $\boldsymbol{\alpha}_{id}$, $\boldsymbol{q}$ and $\boldsymbol{t}$ respectively.

Given an input image $\mathcal{I}$, the three outputs of our model can be expressed as: $\boldsymbol{\alpha}_{id} = \mathcal{S}(\mathcal{E}(\mathcal{I}))$, $\boldsymbol{q} = \mathcal{Q}(\mathcal{E}(\mathcal{I}))$ and $\boldsymbol{t} = \mathcal{T}(\mathcal{E}(\mathcal{I}))$. For better initial conditions, we initialize the output layers of the three MLP in order to predict $\hat{\boldsymbol{s}} = \vec{0}$, $\hat{\boldsymbol{q}} = [1, 0, 0, 0]$ and $\hat{\boldsymbol{t}} = [0, 0, -60]$, values that project the mean 3D shape to the center of the image. All the models are trained using Adam [19] with a learning rate of 10^{-4}, batch size of 32 samples and a total of 60 epochs.

We train our models using a private dataset formed by more than 6,000 3D scans of different subjects acquired using Structure Sensor. Each entry of the database represents a subject, and it is composed by a 3D scan of the face with neutral expression, an average of four images of the subject from multiple views, and the intrinsics and extrinsics matrices of the cameras associated to each image.

We register a 3D template to the 3D scans using a Non-Rigid ICP algorithm in order to work with a fixed topology. Then, a Procrustes analysis is performed with all the registered models, and PCA is applied to extract the identity bases Φ_{id} and the associated eigenvalues Λ. Finally, we express each model on its PCA basis in order to obtain the labels for training. For data augmentation purposes, each database entry is fully symmetrized, computing the symmetric 3D models, and its associated symmetric images and respective symmetric cameras. After the symmetrization, the dataset contains nearly 50,000 images and 12,000 scans.

4.2. Multiterm vs Single term

Our first experiment is performed to validate the hypothesis that MRL not only reduces complexity during training, but also can lead to better results than using a classic multiterm loss.

We define a standard multiterm loss composed of four terms: one for each of the three outputs of our model $\boldsymbol{\alpha}_{id}$, $\boldsymbol{q}$ and $\boldsymbol{t}$, and one for regularization of $\boldsymbol{\alpha}_{id}$:

$$\mathcal{L} = ||\Phi(\boldsymbol{\alpha}_{id} - \hat{\boldsymbol{\alpha}}_{id})||_1 + \alpha||\boldsymbol{q} - \hat{\boldsymbol{q}}||_1 + \beta||\boldsymbol{t} - \hat{\boldsymbol{t}}||_1 + \gamma||\boldsymbol{\alpha}_{id}||_2^2, \tag{12}$$

where α β and γ are weighting hyperparameters. For the identity parameters $\boldsymbol{\alpha}_{id}$, we select the loss proposed by [22], which computes the error of the 3D shape instead of the error of the shape parameters, since it is reported to provide better results.

A random search is performed across α, β, γ hyperparameters. We sample 20 different combinations from the following discrete sets: $\alpha \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, $\beta \in \{1, 10^{-1}, 10^{-2}, 10^{-3}\}$ and $\gamma \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$.

Table 1 shows the best results obtained in terms of camera pose (first row) and 3D shape (second row) across all the considered combinations of α, β, γ. The third row of the table shows the performance achieved with a single training with MRL.

Results indicate a superior performance of the model trained with MRL in terms of reprojection error 1, which is the most representative metric for 3D reconstruction since it combines both 3D shape and camera pose errors. In addition, MRL also obtains competitive results in terms of 3D shape error and camera pose, despite not having been optimized for them, as in the multiterm models of the first and second rows.

We illustrate why using the Euclidean distance to compute the error of $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{t}}$ is not as reliable as using the reprojection error 5. In figure 4, it can be observed that smaller errors in translation and rotation of the camera pose do not imply smaller errors of the 3D scene reprojected into the image plane. This is due to the fact that errors in translation can be compensated by errors in rotation, improving the resulting projective transformation $P = K[R(\boldsymbol{q})|\boldsymbol{t}]$. It can be observed that the overall alignment is better with a single train of our loss.

The whole optimization process for the multiterm loss takes 8 days using a GeForce GTX 1080 Ti graphics card. On the other hand, the training of the model with the proposed MRL is achieved in only 24 hours with the same resources, which is 8 times faster.

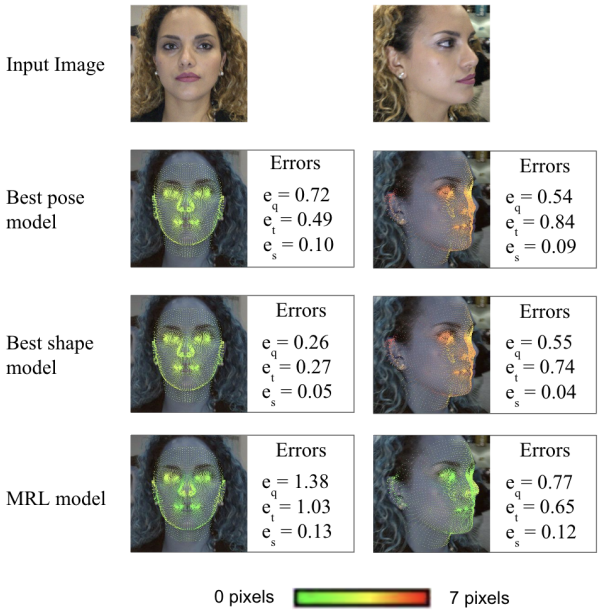

Figure 4: Reprojection errors in image domain produced by the multiterm loss and the MRL.

	Reprojection (pixels)	Shape 3D (mm)	Camera position (mm)	Camera rotation (degrees)
Best pose	5.4	1.7	**2.39**	**2.51**
Best shape	8.58	**1.5**	2.50	2.62
MRL (ours)	**3.29**	1.7	2.86	3.04

Table 1: Prediction errors of the best models for shape and camera pose estimation, compared to the model trained with MRL.

4.3. Robustness against diversity

In order to assess the range of predictions that our model trained using MRL is capable of generating, we randomly select a set of face images with variability in ages, genders, ethnic groups, facial hair and facial accessories.

Qualitatively, Figure 5 shows how our architecture effectively generates accurate 3D reconstructions with high similarity to their respective images under several different facial attributes. Moreover, it is robust against a wide range of light conditions and environments.

4.4. Results on FaceWarehouse

In order to test the generalization of our MRL to other datasets and provide results comparable to other works, a qualitative evaluation is performed on FaceWarehouse dataset [4]. It contains range scans from a total of 150 individuals on neutral and 19 different expressions.

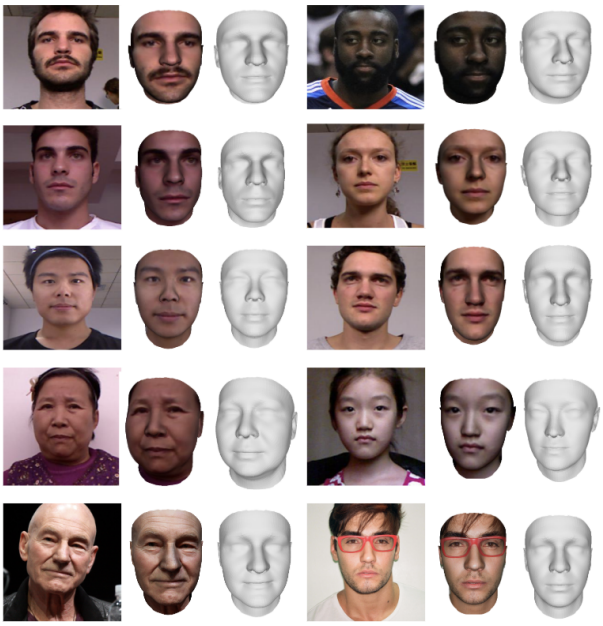

Figure 5: Generalization against face diversity of the architecture trained using MRL.

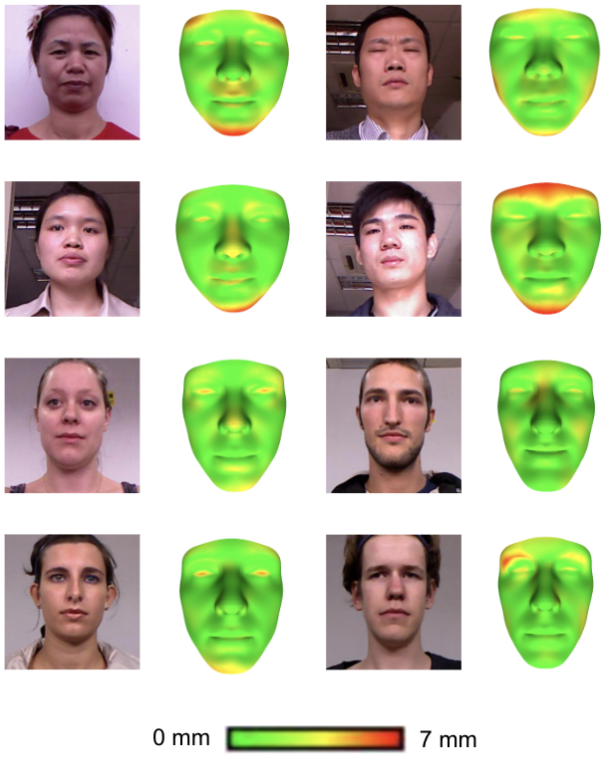

Figure 6: Qualitative evaluation of 3D reconstruction accuracy applied to different subjects in neutral expression from FaceWarehouse dataset [4].

For each image in the dataset containing a case in neutral expression, we infer 50 identity parameters using our architecture trained with the proposed MRL and we convert them into a 3D shape using the 3DMM. Then, we use the Iterative Closes Point algorithm to align the predicted and groundtruth models and compute the point to surface Euclidean distance.

The proposed method effectively estimates the 3D shape of subjects from different gender, ages and ethnicities from a single image. As it can be observed in Figure 6 we obtain low errors that are comparable to the state of the art methods [29, 23].

4.5. Results on MICC dataset

Finally, we evaluate the architecture trained with the MRL on the MICC Florence Faces dataset [1]. It is formed by a set of 3D face scans from 53 different subjects in neutral facial expression. The images are provided in videos recorded under controlled and uncontrolled environments, and in indoor and outdoor scenes, fact that allows evaluating 3D reconstruction algorithms under different levels of complexity.

Single view 3D estimation is performed on the most frontal frame of each individual as in [30], which is always contained in the indoor controlled subset. We align the resulting 3D reconstruction to the groundtruth following the procedure described in Section 4.4. For fair comparison with previous methods, instead of computing the Euclidean distance for each predicted vertex to the surface, we adopt the 3D Root Mean Square Error (3DRMSE). Hence, given a predicted 3D shape $\hat{\boldsymbol{x}}$ and its associated groundtruth $\boldsymbol{x}$, the metric is defined as $3DRMSE = \sqrt{\sum_i (\boldsymbol{x} - \hat{\boldsymbol{x}})^2 / N_v}$, where N_v is the number of vertices.

Method	3DRMSE
3DMM [24]	$1.75 \pm .42$
Flow-based [11]	$1.83 \pm .39$
Discriminative [30]	$1.57 \pm .33$
MRL (ours)	$\mathbf{1.47 \pm .30}$

Table 2: 3DRMSE of different methods evaluated on MICC dataset using a single view.

We compare our method against the state of the art work [30], and previous methods Flow-based [11] and 3DMM [24]. As shown in Table 2, we obtain slightly better results than all previous single view methods without the need of fine tuning any hyperparameter related to the loss.

To end with, we also provide qualitative results in Figure 7 under controlled and uncontrolled environments. Given

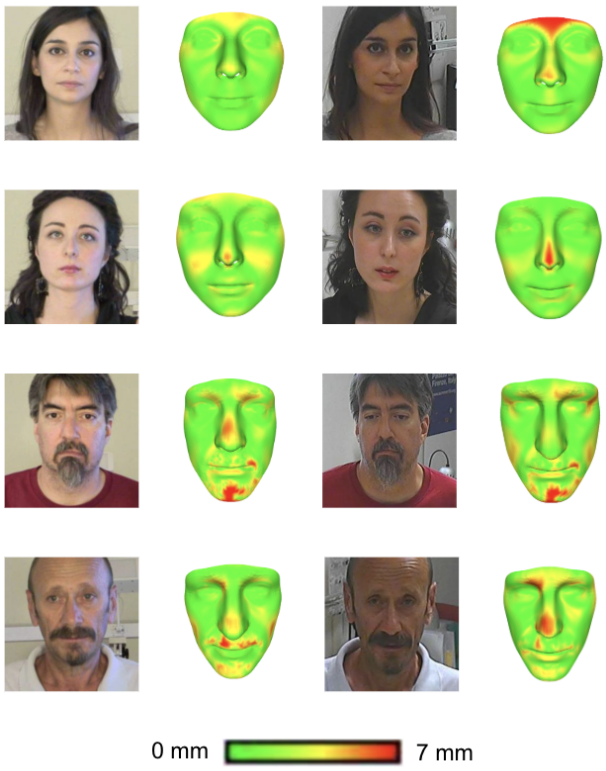

Figure 7: Qualitative results on MICC dataset under controlled (left) and uncontrolled (right) indoor environments

that groundtruth 3D scans from MICC dataset are obtained with the subjects in neutral expression, the uncontrolled environment provides images misaligned with the 3D data causing higher errors.

5. Conclusions

In this work, we proposed a geometry based loss called MRL for learning single view 3D reconstruction problems that does not contain any hyperparameters. Using our approach, the amount of time invested for training deep learning models can be drastically reduced with respect other multiterm losses. Moreover, we proved empirically how models trained using the MRL can simultaneously learn 3D shape and camera pose within a single expression by using random projections into the image planes. We obtain quantitative state of the art results with a single optimization and predict precise 3D shapes under a wide range of different facial characteristics, illuminations and environments.

Although our formulation generalizes to 3DMM with expressions, we only performed tests on 3DMM based on identity basis. The rest of scenarios should be revisited in future work in order to validate the MRL under data presenting more variability. Another limitation of the presented method is the need of real or synthetic data for training. Defining self-supervised losses in the image domain together with multiview supervision could be explored to remove this constrain while maintaining the simplicity of the MRL.

Acknowledgements

This work has been funded by the Industrial Doctoral Programme 2017-DI-028 funded by the Government of Catalonia (Generalitat de Catalunya) through its AGAUR office. This research was partially supported by the Spanish Ministry of Economy and Competitivity and the European Regional Development Fund under contract TEC2016-75976-R (MINECO/FEDER, UE). We acknowledge the support of NVIDIA Corporation for the donation of GPUs.

References

[1] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80. ACM, 2011. 2, 5, 7

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 2, 3

[3] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 2, 3

[4] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2, 5, 6, 7

[5] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 3

[6] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3

[7] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016. 1, 2

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2

[9] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 1, 2

[10] C. Häne, S. Tulsiani, and J. Malik. Hierarchical surface prediction for 3d object reconstruction. *arXiv preprint arXiv:1704.00710*, 2017. 2, 3

[11] T. Hassner. Viewing real-world faces in 3d. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3607–3614, 2013. 7

[12] B. K. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970. 1, 2

[13] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1031–1039. IEEE, 2017. 2, 3

[14] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. *arXiv preprint arXiv:1712.06584*, 2017. 3

[15] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. Learning category-specific mesh reconstruction from image collections. *arXiv preprint arXiv:1803.07549*, 2018. 2, 3

[16] A. Kar, C. Häne, and J. Malik. Learning a multi-view stereo machine. In *Advances in Neural Information Processing Systems*, pages 364–375, 2017. 2, 3

[17] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. *arXiv preprint arXiv:1711.07566*, 2017. 3

[18] A. Kendall, R. Cipolla, et al. Geometric loss functions for camera pose regression with deep learning. In *Proc. CVPR*, volume 3, page 8, 2017. 4

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[22] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 460–469. IEEE, 2016. 2, 3, 5, 6

[23] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5553–5562. IEEE, 2017. 2, 3, 5, 7

[24] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 986–993. IEEE, 2005. 7

[25] J. Roth, Y. Tong, and X. Liu. Unconstrained 3d face reconstruction. *Trans. Graph*, 33(4):43, 2014. 2

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on*, pages 2564–2571. IEEE, 2011. 2

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[28] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. *arXiv preprint arXiv:1712.02859*, 2, 2017. 3, 5

[29] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017. 2, 3, 4, 5, 7

[30] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502. IEEE, 2017. 7

[31] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. *arXiv preprint arXiv:1804.01654*, 2018. 2, 3

[32] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. Reynolds. structure-from-motionphotogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314, 2012. 1

[33] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016. 1, 2, 3

[34] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 3