A Novel 3D Face Reconstruction Model from a Multi-Image 2D Set

Mohamed Dhouioui[1,*], Tarek Frikha[1], Hassen Drira[2] and Mohamed Abid[1]
[1]*CES-Lab, ENIS, University of Sfax, Sfax, Tunisia*
[2]*Centre e Recherche en Informatique Signal et Automatique de Lille, IMT Lille Douai University, Lille, France*

Keywords: Facial Reconstruction, 3D Morphable Model, 3D Face Imaging, Multi-Image 3D Reconstruction, Single-Image 3D Reconstruction.

Abstract: Recently, many researchers have focused on 3D face analysis and its applications, and put a lot of work on developing its methods. Even though 3D facial images provide a better representation of the face in terms of accuracy, they are harder to acquire than 2D pictures. This is why, wide efforts have been put to develop systems which reconstruct 3D face models from 2D images. However, the 2D to 3D face reconstruction problem is still not very advanced, it is both computationally intensive and needs great space exploration to acquire accurate representations. In this paper, we present a 3D multi-image face reconstruction method built over a single image reconstruction model. We propose a novel 3D face re-construction approach based on two levels, first, the use of a single image 3d re-construction CNN model to represent vectorial embeddings and generate a 3d Face morphable model. And second, an unsupervised K-means model on top of the single image reconstruction CNN Model to optimize its results by incorporating a multi-image reconstruction. Thanks to the introduction of a hybrid loss function, we are able to train the model without ground truth reference. Further-more, to our knowledge this is the first use of an unsupervised model alongside a weakly supervised one reaching such performance. Experiments show that our approach outperforms its counterparts in the literature both in single-image and multi-image reconstruction, and it proves that its unique and original nature are very promising to implement in other applications.

1 INTRODUCTION

Facial analysis is widely used in many different applications, we could cite for example interactions between humans and computers (Zhang et al., 2013), security applications (Kaplan et al., 2015), (Burton et al., 1999), motion pictures (Weise et al., 2011), (Weise et al., 2009), and health (EL Rai et al., 2015), (Suttie et al., 2013). In recent years, incorporating 3D data is becoming a trend to surpass some of the inherent issues of the vastly studied 2D facial analysis. A 2D image is insufficient to precisely represent the geometry and full data of a face due to its 3D nature, since it collapses one of the dimensions. Moreover, 3D imaging yields a representation of the facial shape that preserves more or less illumination and pose, two of the primary drawbacks of 2D imaging. As a result, the benefits provided by 3d face recognition techniques come at the expense of a more sophisticated imaging procedure., which is more demanding in data collection and exploration. Some of the well-known techniques for 3D facial information acquisition is stereo-vision systems (Alexander et al., 2010), (Beeler et al., 2011), 3D laser scanners (Lee et al., 1995) (e.g. NextEngine and Cyberware), and RGB-D cameras (such as Kinect). But, each of the mentioned tech-niques has its own drawbacks. Stereo-vision and laser take precise facial scanning, but need controlled settings and costly equipment. As opposed to RGB-D cameras which are cheaper and easier to use, but provide scans with limited quality (Yang et al., 2015).

Researchers propose a substitute approach to acquiring 3D facial scans, they propose to predict its shape from an uncalibrated image (Booth et al., 2018), (Guo et al., 2019). This approach to reconstruct 3D models from 2D images aims to incorporate the ease of 2D picture capture with the advantages of representing the face in a 3D geometry. Despite its attractive-ness, it is inherently ill posed: each and every one of the individual facial geometries, The head's position and texture (including lighting and color) must be retrieved from a single image., which leads to a much more complex problem. As a consequence, and due to the difficulty in determining

Dhouioui, M., Frikha, T., Drira, H. and Abid, M.
A Novel 3D Face Reconstruction Model from a Multi-Image 2D Set.
DOI: 10.5220/0011749500003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 745-753
ISBN: 978-989-758-634-7; ISSN: 2184-4321

which of several 3D faces that make up a single 2D image is the one that best represents the underlying geometry, there may be ambiguities in the solution of the 3D from 2D face reconstruction. Recent research progress has helped to achieve remarkably convincing reconstructions based on newly proposed methodologies, enabling 3D from 2D face reconstruction to be used for number of disciplines (Tu et al., 2019), (Zhu et al., 2019).

The use of past information to resolve ambiguities in the solutions is critical to assure best conditions for 3d from 2d reconstruction approaches. Through literature we could identify in the last decade three techniques for incorporating this previous knowledge: statistical model fitting, photometric stereo, and Deep learning.

A starting method builds a 3D facial model from a set of 3D facial scans, and then encoded prior knowledge into it which is later fitted to the input images.

In another method, they often combine data from many photos, which therefore adds to the complexity. The facial surface normals are estimated using photometric stereo methods in conjunction with 3D face template or 3D facial models.

In the third approach, the 2D to 3D correspondence is implemented using neural net-work models. Given adequate training images, these networks can understand the descriptors required to connect the shape and look of faces.

2 RELATED WORKS

As stated earlier in the introduction deep learning techniques encode past information in the trained network's weights and effectively understand mappings between the 2D picture and the face model. And even though deep learning has proven to be highly useful in several applications, using it directly in 3D from 2D facial reconstruction is constrained by the absence of 3D facial scans that serve as a reliable source of data. To overcome this problem of a lack of ground truth data, academics have suggested many methods for creating and learning from realistic representative training data.

In this section, we present the most relevant works in 3D-from-2D face reconstruction using deep learning as the main tool. Many elements are involved in the learning process. To simplify the study and keep it related to the approach we propose, we will only consider two representative ones in our literature review both the learning framework and the training set that were utilized to train the network. And, according to these elements, we have organized this section.

2.1 Training Data Set

The absence of ground truth data to be utilized as training data is, as we indicated above, the main challenge when using deep learning to 3D-from-2D face reconstruction. This is due to the complexity and hardship of obtaining a huge number of 3D facial scans together with their corresponding 2D pictures. To surpass this limitation, researchers proposed methods either for using 2d datasets and then compare the results to approaches based on 3D datasets or simply constructing artificial training sets and use 3DMMs that have already been constructed to more easily generate 3D faces.

To create synthetic training sets, three basic methodologies can be distinguished. Fit&Render is how we ascribe to the first method.: It entails fitting a 3DMM to real-world photos and then producing synthetic images with the predicted 3D faces. The second one, Generate&Render: involves creating 3D faces by randomly sampling from a 3DMM and then rendering synthetic pictures with the resulting 3D faces. On the other hand, in recent years, a novel method has emerged that involves self-supervised training, which eliminates the necessity for matched 2D and 3D data, therefore the need for artificial data creation.

(Zhu et al., 2016) first proposed the Fit&Render approach, which uses a face profile technique to create pictures for bigger poses and build the 300W-LP (300W large poses) database. They began by fitting a 3DMM using (Romdhani & Vetter, 2005) and (Xiangyu Zhu et al., 2015) over the back-drop to estimate a 3D mesh over the provided facial picture. Next, the 3D mesh was projected over the image and rotated to create a synthetic version of the original image with a bigger posture. Many other writers have used this 300W-LP database (Bhagavatula et al., 2017), (Feng et al., 2018) since it has complicated and realistic face images., as well as 3DMM and projection settings.

The Generate&Render approach involves taking samples from a 3DMM to assemble real-world 3D faces and then generating matching 2D pictures by yielding the 3D faces under various circumstances. (Poses, lighting, etc.). Because this technique does not require a separate 3DMM fitting algorithm, the reconstructive procedure does not limit the network's capacity for learning. 2D pictures, however, are not realistic, as opposed to the Fit&Render method, because the rendering process is completely

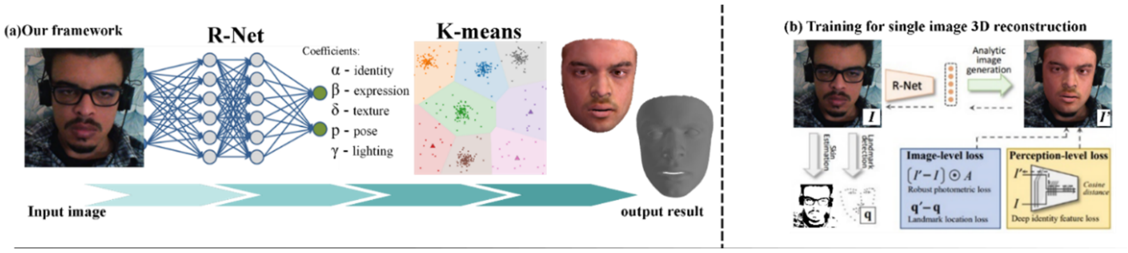

Figure 1: Overview of our approach. (a) The full pipeline of reconstruction using both models (b) The training pipeline of R-Net using the hybrid loss function.

syn-thetic, with synthetic backdrops, lighting settings, projection parameters, and so on. Furthermore, this strategy does not get beyond the drawback of learning from linearly-modeled data because a 3DMM is still needed to build the ground truth 3D faces.

The works that employ the Generate&Render methodology provide a number of solutions to deal with these drawbacks, including the incorporation of real data, the addition of artificial deformations to 3D faces, and the use of more intricate training frameworks.

With the scarcity of realistic ground truth 3D to 2D coupled data and the limitations of using synthetic sets, self-supervision, an innovative method, gained popularity. The crucial idea is that teaching is provided by the data itself by introducing a render-er layer at the network's edge (Tewari et al., 2017), (Tewari et al., 2018), and (Shang et al., 2020).

As a result, using the sampling probabilities discovered from photos of underlying data, the network forecasts attributes for photos without reference data.

2.2 Learning Framework

The core of the deep learning method is the network itself, namely how it is designed and how it gets to know its parameters. A single neural network trained in a single pass is the simplest basic learning framework. Alternatively, other researchers trained their networks iteratively and/or took advantage of the possibilities of more sophisticated designs composed of several networks, such as encoder-decoder architectures or generative adversarial networks. In addition, some of the authors taught each of the many networks to execute certain subtasks. Despite the fact that triangular meshes are the most common means of expressing 3D face data, most works rely on alternative representations, such as depth maps and 3DMM parameters. This is ow-ing to the difficulty of traditional 2D convolution-based networks to interpret non-Euclidean input like meshes. However, a current study topic known as geometric deep learning investigates ways to extend convolutional networks to non-Euclidean inputs, allowing for direct dealing with 3D face models.

When working with 2D data, convnets (CNNs) have shown impressive results, prompting researchers to employ CNNs to reconstruct the 3D face from uncalibrated 2D photos (Tewari et al.,2017), (Tewari et al., 2018), (Shang et al., 2020), and (Wu et al., 2019). (Tewari et al., 2017), (Tewari et al., 2018), (Shang et al., 2020) used the AlexNet (Misra et al., 2016) to teach a CNN to learn from a single image both the renderer (perception and illumination characteristics) and 3DMM configurations (form, ex-pressions, and texturing). At first, the resultant reconstructions in (Tewari et al., 2017) work were crude, and face features were not preserved. They, later (Tewari et al., 2018), (Shang et al., 2020) improved the coarse face produced by using a model that was learned from the training data, AlexNet calculates relative motion for every polygon as coefficients

Contrarily, (Wu et al., 2019) and (Ramon et al., 2019) used the CNN to retrieve information from the image, which were then processed by numerous adjacent layers to individually reconstruct the 3DMM and projection settings. Since the VGG-Net is a deeper network than the AlexNet (Misra et al., 2016), they used it as a feature extractor because it enables them to harvest more important properties even though it is slow.

(Tran et al., 2018) used numerous perspectives in the same way as (Wu et al., 2019), and (Ramon et al., 2019) did. The former sought face recognition and therefore taught the ResNet to be discriminative by utilizing a training set containing the same 3D face related to several images of the same individual. (Deng et al., 2019) retrieved the final reconstruction by linearly integrating single-image reconstructions based on confidence ratings calculated by a second network. As a result, a more precise reconstruction contributed more to the final re-construction, yielding better outcomes than averaging the forms. Unlike (Deng et al., 2019), (Shang et al., 2020) used different perspectives to further push the reconstruction process. They used Multiview consistency to rebuild the 3D face matching to a target image from two nearby views, allowing them to tweak the final reconstruction using all three images at the same time. Nonetheless, this technique requires the authors to train their network using adjacent images, whereas (Deng et al., 2019) could train their networks with a collection of images that were "unrelated."

At contrary to the single pass training used by the works we just discussed, some writers advocated training their networks iteratively. We may differentiate two approaches: one is focused on iteratively enhancing the synthetic training set, while the other is based on repeatedly refining the previous iteration's outcome. The second technique is analogous to a cascaded regressor in that each regressor estimates an update of the input parameters estimated by the preceding regressor, bringing them closer to the ground truth. The majority of the planned works used the same architecture across all iterations (Richardson et al. 2016), (Sanyal et al. 2019).

A ResNet-based network was proposed by (Richardson et al. 2016) and (Sanyal et al. 2019). For the first team the pose is pre-computed using (Kazemi & Sullivan, 2014) and they trained the ResNet to estimate the 3DMM parameters. (Sanyal et al. 2019) estimated the pose in conjunction with the 3DMM parameters. They trained their network by making use of several perspectives and increasing the shape distance between parameters of different persons while decreasing it between parameters of the same person.

3 MATERIALS AND METHODS

Fig. 1 (a) shows how we use a convolutional neural network to regress the coefficients of a 3DMM face model. We additionally regress the lighting and facial pose for unsupervised/weakly supervised training (Tewari et al., 2017), (Tewari et al., 2018) to enable analytic image regeneration. Below a description is detailed of how our model works and outputs in further depth, namely using its three mathematical components; a 3D Face Model, an illumination model, and a camera model.

An affine system of equations can clearly describe how the shape of a face, noted S, and its corresponding texture, noted T, can be represented:

$$S = S(\alpha, \beta) = \bar{S} + B_{id}\alpha + B_{exp}\beta$$
$$T = T(\delta) = \bar{T} + B_t\delta$$

$\bar{T}$ and $\bar{S}$ represent the average values of T and S. We scale with a standard deviation the PCA bases of identity noted B_{id}, B_{exp}, and B_t.

α, β, and δ represent the coefficient vectors for generating a 3D face. The well-known Basel Face Model from (Paysan et al., 2009) is used to determine $\bar{S}$, B_{id}, $\bar{T}$ and B_t. And from (Guo et al., 2019) we use the expression bases B_{exp} which is built from FaceWarehouse (Chen Cao et al., 2014). By excluding the neck and ear regions an selecting a subset of the bases resulting in $\alpha \in R^{80}$, $\beta \in R^{64}$ and $\delta \in R^{80}$ we can get our resulting model that contains 36K vertices.

A Lambertian surface for face is assumed. And the scene illumination is approximated with Spherical Harmonics (SH) (Ramamoorthi et al. 2001). Given the following equation:

$$C(n_i t_i|\gamma) = t_i \sum_{b=1}^{B^2} \gamma_b \varphi_b(n_i)$$

We can calculate S_i being the radiosity of a vertex, n_i being the surface normal, and t_i being the skin texture. Spherical harmonics basis functions are noted $\varphi_b: \mathbb{R}^3 \to \mathbb{R}$ and their corresponding coefficients are noted γ_b. Just like (Tewari et al., 2017),(Tewari et al., 2018), the number of bands is chosen as B=3 bands and monochromatic lights are assumed such that $\gamma \in \mathbb{R}^9$.

For the geometry of the 3D-2D projection, we employ the perspective camera model with an experimentally chosen focal length. A translation **t** and a rotation **R** are used to represent the 3D face pose p such as $\boldsymbol{R} \in \boldsymbol{SO}(\boldsymbol{3})$ and $\boldsymbol{t} \in \mathbb{R}^{\boldsymbol{3}}$. The output of our model is a vector representing the unknowns to be determined$x = (\alpha, \beta, \delta, \gamma, p) \in \mathbb{R}^{\boldsymbol{239}}$. In this paper, by modifying the last fully-connected layer to 239 neurons, we try to regress the mentioned coefficients using a ResNet-50 neural network (He et al., 2016) (Deng et al., 2019). Henceforth, we will be referring to this ResNet-50 model as R-Net. In the following sections, we present how we train it.

3.1 Single Image Reconstruction

As stated previously, R-net is used to regress a coefficient vector x as an output, and for this task the model's input is an RGB image noted **I**. When passing the image **I**, we get its corresponding vector **x** as output, this vector is then used to analytically generate a reconstructed image **I'** (Fig. 1. shows some examples of this process). By backpropagating a hybrid-level loss evaluated on **I'**, we can train R-net without any ground truth labels.

This hybrid-level loss is composed of two functions; an Image-level loss, and a Perception-level loss.

As for the image level loss, we used the same skin-aware photometric loss as (Deng et al., 2019) such as:

$$L_{photo}(x) = \frac{\sum_{i \in M} A_i . \|I_i - I'_i(x)\|_2}{\sum_{i \in M} A_i}$$

Where:

- i is pixel index.
- M denotes the reprojected face region.
- A is a skin attention mask created using a naïve Bayes classifier on a skin image dataset [26] and for each pixel setting $A_i = \begin{cases} 1 \; if \; P_i > 0.5 \\ P_i \; otherwise \end{cases}$ and as seen bellow Fig. 2. shows the results with (bottom row) and without (top row) a skin attention mask.

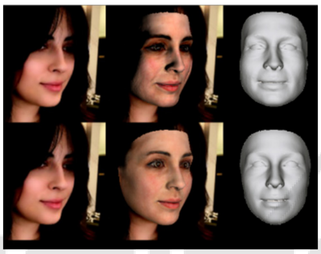

Figure 2: The effect of using a skin attention mask.

(Top row: without mask, bottom row with mask) (Deng et al., 2019)

And for face alignment, we use the method of (Suttie et al., 2013) in detecting 68 facial landmarks {qn} for the training images. By projecting the landmarks vertices of our reconstructed shape onto the resulting image I' we obtain {q'n} which we use to compute the loss as:

$$L_{lan}(x) = \frac{1}{N} \sum_{n=1}^{N} \omega_n \left\| q_n - q'_n(x) \right\|^2$$

Experimenting with ω_n led to setting it to 20 for mouth and nose points and 1 for others.

As for the perception-level loss, inspired by (Yang et al., 2015) and (Deng et al., 2019) we used the following cosine distance:

$$L_{per}(x) = 1 - \frac{< f(I), f\big(I'(x)\big) >}{\|f(I)\| . \|f(I'(x))\|}$$

With f(.) and <.,.> representing correspondently deep feature encodings and vector inner product.

In our work, we use as deep feature extractor a pretrained FaceNet structure (Schroff et al., 2015).

The regressed 3DMM coefficients could contain some shape or texture degeneration on the face, in order to prevent this from happening we apply another loss function on said coefficients to impose a distribution respecting the mean face:

$$L_{coef}(x) = \omega_\alpha \|\alpha\|^2 + \omega_\beta \|\beta\|^2 + \omega_\gamma \|\gamma\|^2$$

With:

$$\omega_\alpha = 1.0, \qquad \omega_\beta = 0.8, \qquad \omega_\gamma = 1.7e - 3$$

The Basel 2009 3DMM contain some dried shading and we would like to maintain a constant skin shading like (Tewari et al., 2018), for this we add a constraint that reprimands the texture map variance:

$$L_{tex}(x) = \sum_{c \in \{r,g,b\}} var(T_{c,R}(x))$$

With R being a predefined skin region involving the forehead, cheek, and nose.

To summarize our training loss L(x) could be described as a composition of three levels; a first level having two image-level losses L_{photo} and L_{lan}, a second level having one perceptual loss L_{per}, and a third level having two regularization losses L_{coef} and L_{tex} . The corresponding weights for these losses are set to the following values throughout our experiment:

$$\omega_{photo} = 1.9, \quad \omega_{lan} = 1.6e - 3, \quad \omega_{per} = 0.2, \quad \omega_{coef} = 3e - 4, \quad \omega_{tex} = 5$$

3.2 Multi-Image Reconstruction

Even though reconstructing a face from a single image input seems to be a great endeavour, such reconstruction could be somewhat lacking in terms of precision and resolution. As we made clear from literature research, having multiple images of a face would affect greatly the performance of a model in its reconstruction task, since single images could be subject to bad lighting or occlusions.

In this section of the paper, we propose building an unsupervised machine learning model that goes hands in hands with our previously created R-Net. This model would make use of the output obtained from R-Net and create a spacial representation of different images of the same object, thus gaining more information about the face and resulting in better metrics for the reconstruction.

Creating and training a model with an arbitrary number of images representing the same entity is not a straightforward task. In this work we use K-means as an aggregation algorithm to search for a vector representation within a cluster of vectors obtained from R-Net.

Given a set of M subjects with each having j images, our approach could be described as follows;

- For each subject, we start by generating the reconstructed image set $\{I^{j'}\}$ of $\{I^{j}\}$. Thus, resulting in having $x^j = (\alpha^j, \beta^j, \delta^j, p^j, \gamma^j)$ the output vector of R-Net for each image **j**.
- After obtaining $M \times j$ vectors, we create a K-means models with K = M the number of clusters.
- Shuffle the vectors dataset and initialize the centroids.
- Then, we keep iterating until centroids no longer change, meaning that the assignment of data points to individual clusters isn't changing anymore.
- By now, the algorithm computes the sum of the squared distance between centroids and all data points.
- After computing distance equations, our model assigns each data point to its closest cluster.
- Finally, the centroids for each cluster are computed by taking the average of the data points that belong to each cluster
- The resulting clusters correspond each to a vector $x_i = (\alpha_i, \beta_i, \delta_i, p_i, \gamma_i)$ with $i \in [1, M]$. Using these vectors, we reconstruct the face model for each person respectively.

A last thing to note is that given the iterative nature of K-means algorithm and the random initialization of centroids at the start, an issue may arise with different initializations leading to different clusters. Therefore, we recommend using the same approach we did, which is to run the algorithm using different initializations of centroids and picking the run yielding the lower sum of squared distance. Moreover, since the nature of vectors obtained from R-Net correspond to facial identities convergence is achieved more easily since all the individual's feature vectors are close to each other in distance.

4 RESULTS

Training the R-Net model was done using multiple sources namely; CelebA(Liu et al., 2015), 300W-LP(Zhu et al., 2016), I-JBA(Klare et al., 2015), LFW(B. Huang et al., 2008) and LS3D(Bulat & Tzimiropoulos, 2017). Using these images, we took in consideration the balancing of pose and race distributions and got approximately 260K face images.

The input size was set to 224x224. We used the pretrained weights of ImageNetas initialization and then trained the R-Net model using Adam optimizer, a batch size of 8 and starting with a learning rate of 1e-4 ending after 500K iterations.

As for K-means, we used an image set that is composed of our own facial recognition dataset, and 300W-LP(Zhu et al., 2016) from which we choose 5 random images for each person in various poses and lighting. The resulting data-set has approximately 20K images of 5K identities.

Some of the reconstruction results seen in Fig 3. bellow show the resulting images and 3D models of our R-Net model. As observed clearly, the obtained 3D model is very smooth and lacks any visible anomalies.

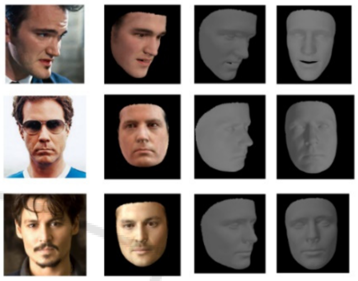

Figure 3: Examples of the results obtained through R-Net model.

For fair comparison with previous results in the current literature, we studied the results of our models on the MICC Florence 3D Face Dataset (Bagdanov et al., 2011) which contains 53 subjects each having a neutral expression ground truth and three video sequences taken in 3 three different scenarios: cooperative, indoor, and outdoor. Table 1 and Fig 4 shows a comparison between the results from our R-Net , the results from (Tran et al., 2017), those from (Genova et al., 2018) and those from (Deng et al., 2019).

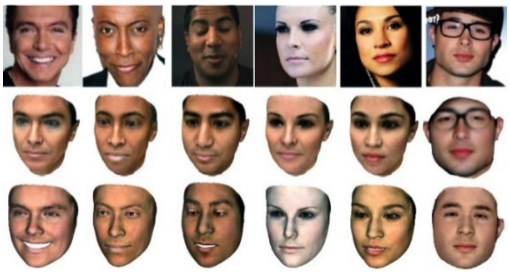

Figure 4: Comparison with the work done by (Genova et al., 2018) in the second row and ours in the last row.

Table 1: Mean Root Mean Squared Error (RMSE) across 53 subjects on MICC dataset (in mm). We use ICP for alignment and compute point-to-plane distance between results and ground truth.

Method	Cooperative	Indoor	Outdoor
(Tran et al., 2017)	1.97±0.49	2.03±0.45	1.93±0.49
(Genova et al., 2018)	1.78±0.54	1.78±0.52	1.76±0.54
(Deng et al., 2019)	1.66±0.52	1.66±0.46	1.69±0.53
Ours	1.65±0.49	1.66±0.57	1.69±0.53

As seen, our results surpass those of (Genova et al., 2018) and (Tran et al., 2017) both in visual renders and RMSE, even though we cut the ground truth meshes to compensate for (Tran et al., 2017) only containing part of the forehead. As for (Deng et al., 2019) our results are very close to each other due to having arguably similar model architecture when it comes to R-Net and working on the level of single image reconstruction.

However, working on multi-image reconstruction, table 2 below shows that using K-means alongside R-Net clearly outperforms other methods.

The fact that our R-Net produces smooth and superior quality face shapes, and that by itself it surpasses some of the literature results, would ensure that adding a second layer on top of it, in our case the K-means, converges for even better results. This, to our knowledge, is the first approach applying K-means to face reconstruction attaining such results. And through qualitative analysis we further demonstrated the interesting contribution done in this paper.

Table 2: Results for multi-image recognition on MICC dataset, using the strategy of (Piotraschke & Blanz, 2016). S and G here denote segment-based aggregation and global based aggregation.

Method	Cooperative	Indoor	Outdoor	All
Shape averaging	1.97±0.49	2.03±0.45	1.93±0.49	1.62±0.51
(Piotraschke & Blanz, 2016)-G	1.78±0.54	1.78±0.52	1.76±0.54	1.65±0.55
(Piotraschke & Blanz, 2016)-S	1.66±0.52	1.66±0.46	1.69±0.53	1.65±0.55
(Deng et al., 2019)	1.60±0.51	1.61±0.44	1.63±0.47	1.56±0.48
Ours	1.59±0.53	1.61±0.50	1.62±0.51	1.54±0.52

5 CONCLUSIONS

In this paper, we designed and trained a ResNet based model for the task of face reconstruction from a single image. This model was trained using a custom loss function that exploits image data levels without the need for 3D ground truth shapes. And it showed great results in comparison with previous work done in the literature review.

Furthermore, we improved the outcomes by using information from multiple image which provides better insights on the facial structure. This was done by taking ad-vantage of the K-means algorithm through its characteristics that computes centroids of data cluster, resulting in far better understanding and representation of feature vectors.

REFERENCES

Zhang, L., Jiang, M., Farid, D., & Hossain, M. (2013, October). Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems With Applications*, *40*(13), 5160–5168.

Kaplan, S., Guvensan, M. A., Yavuz, A. G., & Karalurt, Y. (2015, December). Driver Behavior Analysis for Safe Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *16*(6), 3017–3032.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999, May). Face Recognition in Poor-Quality Video: Evidence From Security Surveillance. *Psychological Science*, *10*(3), 243–248.

Weise, T., Bouaziz, S., Li, H., & Pauly, M. (2011, July). Realtime performance-based facial animation. *ACM Transactions on Graphics*, *30*(4), 1–10.

Weise, T., Li, H., Van Gool, L., & Pauly, M. (2009). Face/Off. *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '09*.

EL Rai, M. C., Werghi, N., Al Muhairi, H., & Alsafar, H. (2015, February). Using facial images for the diagnosis of genetic syndromes: A survey. *2015 International Conference on Communications, Signal Processing, and Their Applications (ICCSPA'15)*.

Suttie, M., Foroud, T., Wetherill, L., Jacobson, J. L., Molteno, C. D., Meintjes, E. M., Hoyme, H. E., Khaole, N., Robinson, L. K., Riley, E. P., Jacobson, S. W., & Hammond, P. (2013, March 1). Facial Dysmorphism

Across the Fetal Alcohol Spectrum. *Pediatrics*, *131*(3), e779–e788.

Alexander, O., Rogers, M., Lambeth, W., Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, & Debevec, P. (2010, July). The Digital Emily Project: Achieving a Photorealistic Digital Actor. *IEEE Computer Graphics and Applications*, *30*(4), 20–31.

Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R. W., & Gross, M. (2011, July). High-quality passive facial performance capture using anchor frames. *ACM Transactions on Graphics*, *30*(4), 1–10.

Lee, Y., Terzopoulos, D., & Walters, K. (1995). Realistic modeling for facial animation. *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '95*.

Yang, L., Zhang, L., Dong, H., Alelaiwi, A., & El Saddik, A. (2015, August). Evaluating and Improving the Depth Accuracy of Kinect for Windows v2. *IEEE Sensors Journal*, *15*(8), 4275–4285.

Booth, J., Roussos, A., Ververas, E., Antonakos, E., Ploumpis, S., Panagakis, Y., & Zafeiriou, S. (2018, November 1). 3D Reconstruction of "In-the-Wild" Faces in Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2638–2652.

Guo, Y., Zhang, J., Cai, J., Jiang, B., & Zheng, J. (2019, June 1). CNN-Based Real-Time Dense Face Reconstruction with Inverse-Rendered Photo-Realistic Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(6), 1294–1307.

Zhu, X., Liu, X., Lei, Z., & Li, S. Z. (2019, January 1). Face Alignment in Full Pose Range: A 3D Total Solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(1), 78–92.

Romdhani, S., & Vetter, T. (n.d.). Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

Xiangyu Zhu, Lei, Z., Junjie Yan, Yi, D., & Li, S. Z. (2015, June). High-fidelity Pose and Expression Normalization for face recognition in the wild. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bhagavatula, C., Zhu, C., Luu, K., & Savvides, M. (2017, October). Faster than Real-Time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Feng, Y., Wu, F., Shao, X., Wang, Y., & Zhou, X. (2018). Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. *Computer Vision – ECCV 2018*, 557–574.

Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016, June). Face Alignment Across Large Poses: A 3D Solution. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Richardson, E., Sela, M., & Kimmel, R. (2016, October). 3D Face Reconstruction by Learning from Synthetic Data. *2016 Fourth International Conference on 3D Vision (3DV)*.

Richardson, E., Sela, M., Or-El, R., & Kimmel, R. (2017, July). Learning Detailed Face Reconstruction from a Single Image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sela, M., Richardson, E., & Kimmel, R. (2017, October). Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Wu, Y., Shah, S. K., & Kakadiaris, I. A. (2016, February). Rendering or normalization? An analysis of the 3D-aided pose-invariant face recognition. *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*.

Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., & Theobalt, C. (2017, October). MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Tewari, A., Zollhofer, M., Garrido, P., Bernard, F., Kim, H., Perez, P., & Theobalt, C. (2018, June). Self-Supervised Multi-level Face Model Learning for Monocular Reconstruction at Over 250 Hz. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tewari, A., Zollhofer, M., Bernard, F., Garrido, P., Kim, H., Perez, P., & Theobalt, C. (2020, February 1). High-Fidelity Monocular Face Reconstruction Based on an Unsupervised Model-Based Face Autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 357–370.

Zhou, Y., Deng, J., Kotsia, I., & Zafeiriou, S. (2019, June). Dense 3D Face Decoding Over 2500FPS: Joint Texture & Shape Convolutional Mesh Decoders. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, F., Bao, L., Chen, Y., Ling, Y., Song, Y., Li, S., Ngan, K. N., & Liu, W. (2019, June). MVF-Net: Multi-View 3D Face Morphable Model Regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sanyal, S., Bolkart, T., Feng, H., & Black, M. J. (2019, June). Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, F., Tran, L., & Liu, X. (2019, October). 3D Face Modeling From Diverse Raw Scan Data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tu, X., Zhao, J., Xie, M., Jiang, Z., Balamurugan, A., Luo, Y., Zhao, Y., He, L., Ma, Z., & Feng, J. (2021). 3D Face Reconstruction From A Single Image Assisted by 2D Face Images in the Wild. *IEEE Transactions on Multimedia*, *23*, 1160–1172.

Misra, I., Shrivastava, A., Gupta, A., & Hebert, M. (2016, June). Cross-Stitch Networks for Multi-task Learning.

2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ramon, E., Escur, J., & Giro-i-Nieto, X. (2019, October). Multi-View 3D Face Reconstruction in the Wild Using Siamese Networks. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*.

Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., & Tong, X. (2019, June). Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., & Quan, L. (2020). Self-Supervised Monocular 3D Face Reconstruction by Occlusion-Aware Multi-view Geometry Consistency. *Computer Vision – ECCV 2020*, 53–70.

Kazemi, V., & Sullivan, J. (2014, June). One millisecond face alignment with an ensemble of regression trees. *2014 IEEE Conference on Computer Vision and Pattern Recognition*.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009, September). A 3D Face Model for Pose and Illumination Invariant Face Recognition. *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, & Kun Zhou. (2014, March). FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, *20*(3), 413–425.

Ramamoorthi, R., & Hanrahan, P. (2001). An efficient representation for irradiance environment maps. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '01*.

Ramamoorthi, R., & Hanrahan, P. (2001). A signal-processing framework for inverse rendering. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '01*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015, June). FaceNet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015, December). Deep Learning Face Attributes in the Wild. *2015 IEEE International Conference on Computer Vision (ICCV)*.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., & Jain, A. K. (2015, June). Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

B. Huang, Ramesh, Berg, & Learned-Miller. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *University of Massachusetts, Amherst*, 07–49.

Bulat, A., & Tzimiropoulos, G. (2017, October). How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). *2017 IEEE International Conference on Computer Vision (ICCV)*.

Bagdanov, A. D., Del Bimbo, A., & Masi, I. (2011, December). The florence 2D/3D hybrid face dataset. *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*.

Tran, A. T., Hassner, T., Masi, I., & Medioni, G. (2017, July). Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., & Freeman, W. T. (2018, June). Unsupervised Training for 3D Morphable Model Regression. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Piotraschke, M., & Blanz, V. (2016, June). Automated 3D Face Reconstruction from Multiple Images Using Quality Measures. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.