Supplementary material to: "Look Ma, no landmarks!" – Unsupervised, model-based dense face alignment

Tatsuro Koizumi[1,2] and William A. P. Smith[2]

[1] Canon Inc.
[2] University of York, York, UK
{tk856,william.smith}@york.ac.uk

In this supplementary document, we provide additional details regarding our choice of 3D morphable model, implementation details and additional experimental results.

A Morphable model details

We employ the Basel Face Model 2017 [6] as a representation of a face, which has $N_s = 199$, $N_e = 100$, and $N_r = 199$ dimensions for facial identity shape, facial expression shape, and skin albedo respectively. We scale basis $\mathbf{s}_j^i$ and $\mathbf{a}_j^i$ so that the standard deviation of α_i and β_i is 1.

Since our differentiable linear least squares layer samples the 3DMM mean and basis for each pixel based on predicted correspondence, we flatten the 3DMM to a 2D parameterisation beforehand. Specifically, we generate a Tutte embedding [4] for each component of the 3DMM. We force the boundary of the embedding to be square. We refer to the flattened 3DMM as UV-3DMM and its domain of definition as UV-space. To fill a hole inside the mouth of the Basel Face Model 2017, we introduce an auxiliary vertex inside the hole and connect it with the boundary vertices of the mouth. We set the mean value of mouth boundary vertices for each component of the auxiliary vertex. The resolution of precomputed UV-3DMM is 320×320 pixels. In our linear least squares layer, we process 3DMM and input data as described in Fig. 10.

B Linear least square layer details

Fig. 10 shows a schematic overview of how the linear least squares solutions for geometric and photometric parameters are combined within our network. The differentiable closed form solution is given in Sec. R.

C Stochastic Sampling

Solving a linear system over all pixels for all images in a minibatch within the network during training is prohibitively computationally expensive. For this reason, we introduce a stochastic sampling of pixels for the linear least square process

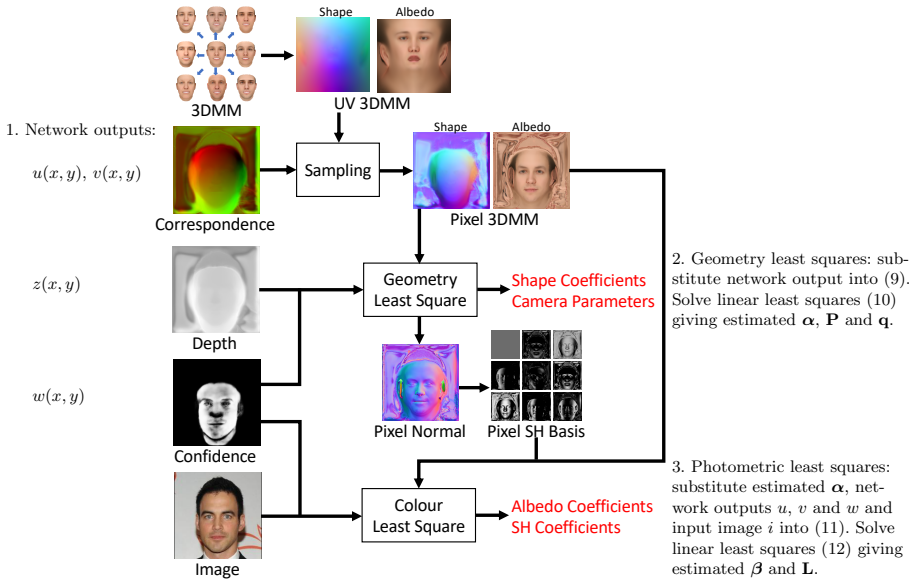
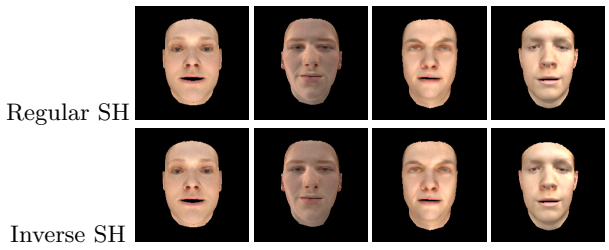


Fig. 10: Overview of linear least square layer (outputs in red). Inputs to this process are the outputs from the image-to-image network.



Max error	0.049	0.109	0.123	0.272
RMS error	0.019	0.043	0.029	0.058

Fig. 11: Empirical validation of inverse spherical harmonic lighting model.

to reduce memory consumption. We randomly select 10,000 pixels which have confidence value larger than $0.001\times$ the maximum confidence value. If the number of pixels which fulfil the above criteria is less than 10,000, we select the rest of the pixels randomly.

D Empirical validation of inverse lighting model

We empirically validate inverse spherical harmonic(SH) lighting model in Fig. 11. The upper row shows randomly generated images based on conventional SH

lighting. We generate random SH coefficients by $\sigma = 0.2$ and add the random lighting to constant lighting with intensity 0.9. We use the same SH coefficients for all RGB channels. The lower row shows images of the same faces rendered based on inverse SH lighting. Inverse SH coefficients are calculated as a least squares solution that minimises the difference between estimated inverse lighting and inverted original lighting at random 100,000 sample points on the sphere. We also show mean and max errors of lighting intensity between conventional and inverse SH lighting model among sample points.

E Stability of photometric least squares

We assume the pixel value in both images and 3DMM is scaled to $[0, 1]$. Dark pixels, with value close to zero, cause numerical unstability so we clamp low pixel values of an input image. Specifically, we apply softplus function to input image as preprocessing: $i_{x,y} = \log(1 + e^{\xi \cdot i_{x,y}})/\xi$ where ξ is a parameter to adjust the scale of softplus function. We also apply inverse function of softplus function to visualise output images. We use $\xi = 4$.

F Network pretraining

We pretrain our network using a small number of roughly aligned images by applying data augmentation by 2D similarity transformation. In pretraining, we directly supervise the pixel-wise prediction network using constant value depth map, synthetic confidence map, and synthetic correspondence map. We aligned mean shape of 3DMM to pretraining images using average 5 landmark position, and generate synthetic confidence map, in which face region is set to 1 and the other to 0, and synthetic correspondence map. The same supervision data is used for all the pretraining images. An example of an input image and supervision data is shown in Fig. 12. We apply random similarity transformation to both input images and supervision data. Though we use roughly aligned image for pretraining, we never use landmarks of each image and 3D ground truth. Thus, our network can be regarded as unsupervised training in the conventional context. We initially pretrain our network using 1k images from pre-aligned CelebA dataset. Here, batch size is 5, number of iterations is 14k.

During early iterations of the main training, we additionally regularise the camera translation parameters in the linear least squares system as the calculation of full perspective camera parameters from planar depth tends to be unstable. Camera translation parameters are regularised by applying L2 distance regularisation between camera viewpoint and a fixed point placed in front of the face.

G Adaptive Loss Adjustment

During training, weights of each 3DMM coefficient in statistical reguralisation E_{stat} is adaptively adjusted so that the exponential average of the squared value

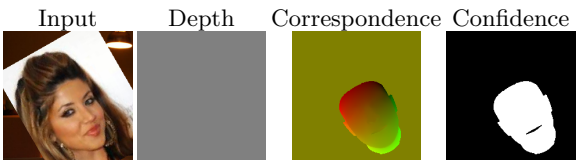


Fig. 12: Example of training data for pretraining.

of each coefficient is kept to be 1, which is equivalent to the variance defined in the 3DMM. The weight of ith coefficient in the jth iteration $\omega_{i,j}$ is set by:

$$\omega_{i,j} = \max(\min(k_{i,j}E[\omega_i]_j, \omega_{\max}\}), \omega_{\min}) \tag{18}$$

where $E[\omega_i]_j = (1-\theta)\omega_{i,j-1} + \theta E[\omega_i]_{j-1}$, $k_{i,j} = \max(\min(E[\alpha_i^2]_j, \alpha^2_{\max}), \alpha^2_{\min})$ and $E[\alpha_i^2]_j = (1-\theta)\alpha^2_{i,j-1} + \theta E[\alpha_i^2]_{j-1}$. Here $k_{i,j}$ represents the rate of change in the jth iteration, and $\alpha_{i,j}$ represents the ith coefficient obtained from outputs in the jth iteration. We clamp both the rate of change and the weight. We set the update ratio as $\theta = 0.05$, and the clamp threshold $\omega_{max} = 10^4, \omega_{min} = 10^{-4}, \alpha^2_{max} = 1.01, \alpha^2_{min} = 0.99$. We initialise the exponential average as $E[\alpha_i^2]_0 = 0.1$ and $E[\omega_i]_0 = 0.1$ before starting training.

H Intermediate output

Fig. 13 shows outputs of the pixel-wise prediction network. Even without the least square 3DMM fitting, the quality of output is also convincing.

I Additional comparison

We also compare our method with the state-of-the-art Deng et al. [3] (Fig. 14). Due to richer supervision based on landmarks and ID, Deng et al. [3] shows better quality. However, our method still has comparable quality despite it is unsupervised method and has robustness against 2D similarity transformation. Fig. 15 shows comparison with Tran et al. [17], MoFA [15], and Genova et al. [5] in 3D visualisation. This indicates our method has comparable quality to other deep learning based 3D face reconstruction methods.

We also evaluate based on the identity of reconstructed faces (Fig. 16,Tab. 1). As Genova et al. [5] optimises facial identity of reconstructed image, it outperforms ours. However, our method is slightly better than Tran et al. [17] and MoFA [15]. This could be a contribution of least squares in colour, which improves fidelity of an output face.

J Additional quantitative results

In this supplementary material, additional quantitative evaluation is provided. We follow the evaluation in Jakab et al. [7] and Thewils et al. [16]. Tab. 2

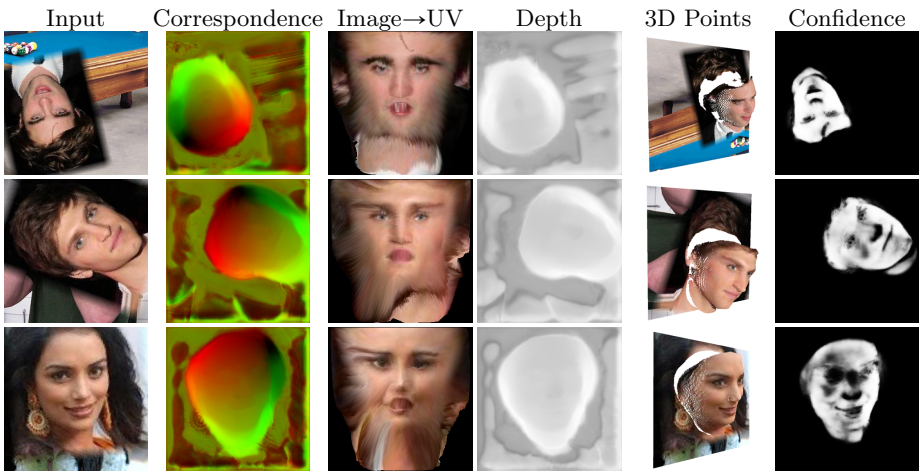


Fig. 13: Intermediate output of pixel-wise prediction network. Left to right: Correspondence map, image mapped to UV-space, depth map, depth map as point cloud and confidence map.

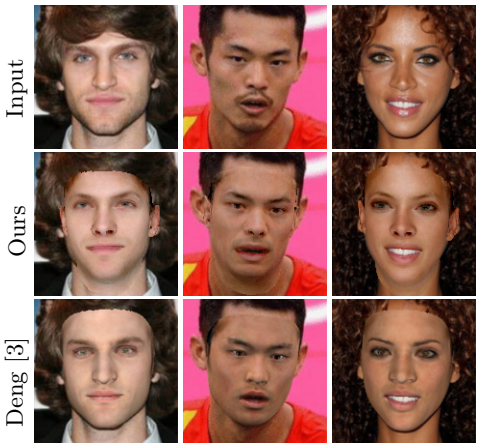


Fig. 14: Comparison to Deng et al. [3].

shows quantitative evaluation in terms of a ratio of standard MSE to the inter-ocular distance expressed as a percentage. In this experiment, the position of five landmarks is evaluated on MAFL test-set [23] and AFLW [11] test-set. Despite our approach is unsupervised, the result is comparable to supervised methods except RCPR [1] and MTCNN [22]. Previous works in unsupervised and self-supervised detection have generally better performance. This might be because those unsupervised/self-supervised methods can exploit indirect supervision from multiple images of a same object as well as a fine-tuned regressor,

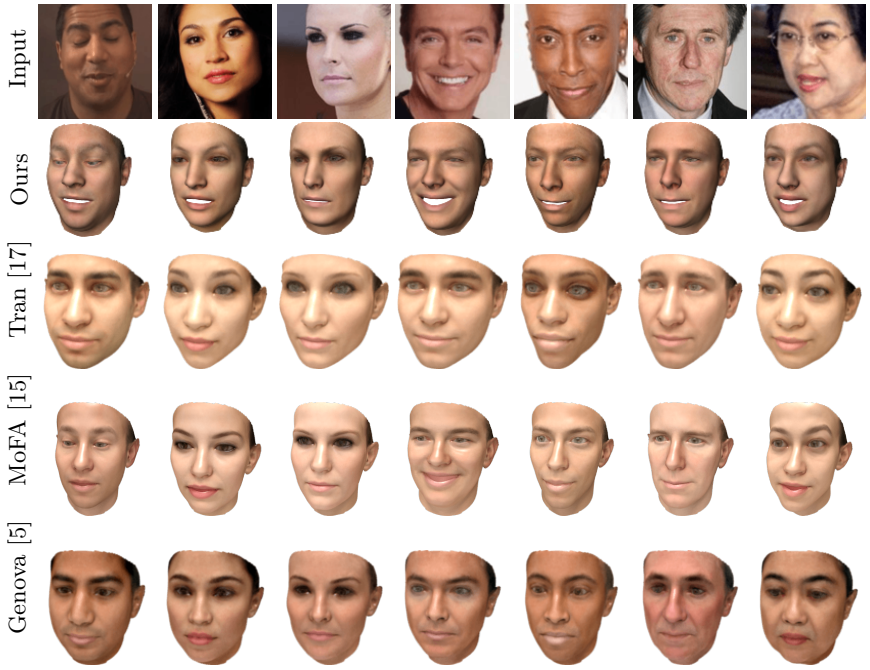


Fig. 15: Reconstructed 3D face from images on MoFA-test dataset. Note that Tran [17] and Genova [5] only reconstruct a neutral face without expression.

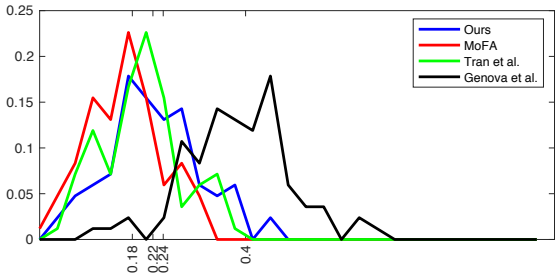


Fig. 16: Distributions of cosine similarity for VGG-Face descriptors between rendering/photo for MoFA-test. Means below.

which converts unorganized landmarks to pre-defined one. In addition, our approach is not as robust as unsupervised/self-supervised methods because our approach is constrained by the linear model and cannot handle outliers such as occlusion and self-shadow properly.

Method	Supervision	Same	Different
MoFA [15]	None/Landmarks	0.30	0.11
Tran et al. [17]	Fully supervised	0.27	0.14
Genova et al. [5]	Identity	**0.09**	**0.32**
Ours	None	0.26	0.16

Table 1: Earth mover's distance between distributions in Fig. 16 and same/different identities on LFW (see [5]). Low distance for "Same" and high distance for "Different" means rendering/photo pair distances are similar to real photo pairs of the same person.

Method	MAFL	AFLW
Supervised		
RCPR [1]	-	11.60
CFAN [20]	15.84	10.94
Cascaded CNN [14]	9.73	8.97
TCDCN [23]	7.95	7.65
RAR [19]	-	7.23
MTCNN [22]	5.39	6.90
Unsupervised/Self-supervised		
Thewils [16]	6.67	10.53
Shu [12]	5.45	-
Zhang [21]	3.16	6.58
Wiles [18]	3.44	-
Jakab [7]	2.54	6.33
Ours		
Direct	8.10	9.82
Fitted	7.86	9.74

Table 2: Quantitative evaluation on MAFL [23] and AFLW [11] Database.

K Additional qualitative results

In this supplemental document, additional results of reconstruction are provided (Fig. 17). We train our network using images from CelebA dataset [9]. For both training and test, we apply random 2D similarity transformation to original cropped CelebA images, and fill the background region with random images from ImageNet [8]. This evaluation shows our network can reconstruct images, which have diverse ethnicity, expression, gender, and pose. Fig. 18 shows reconstruction from rotated and cropped face images in ImageNet. This result indicates the blended boundary of an augmented image have no significant effect for reconstruction.

L Performance versus training iteration

Fig. 19 shows the convergence of reconstructed images during the training. The initial estimate is based on the pretrained network, which only requires small

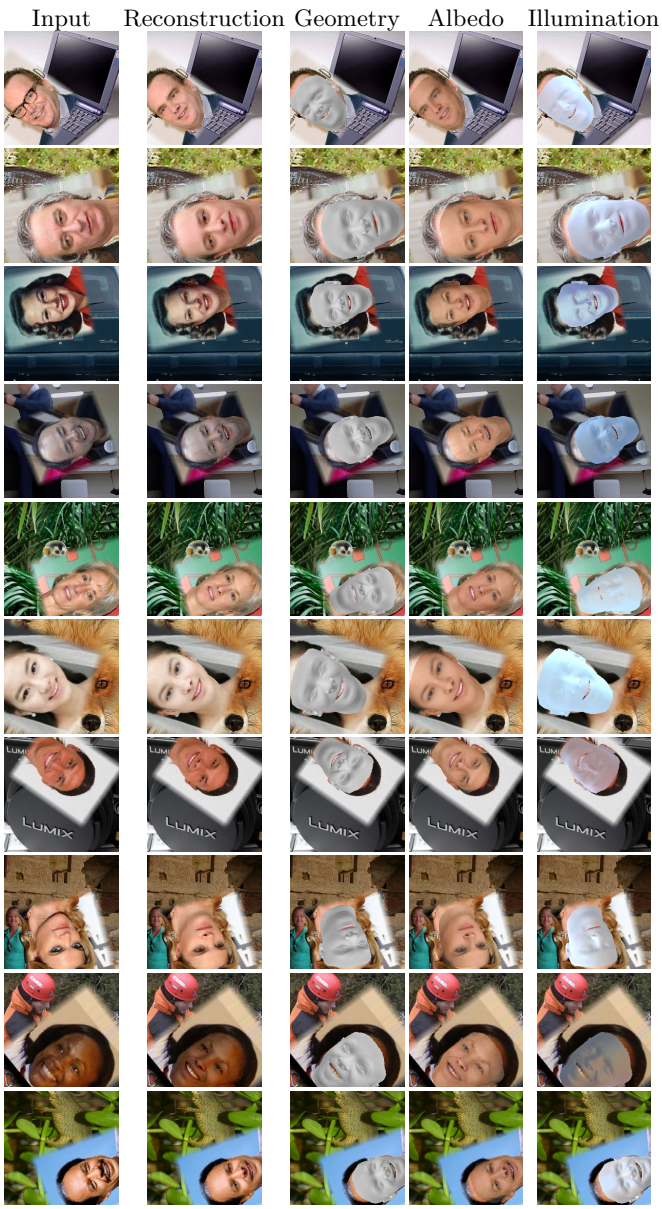


Fig. 17: Reconstruction result from images in CelebA dataset. Random 2D similarity transformation is applied to an original cropped CelebA image, and the background region is filled with a random ImageNet image. Images for this evaluation is not used for the training.

Fig. 18: Reconstruction result from images in ImageNet dataset. Images containing a face are selected and cropped.

amount of roughly aligned images for supervision. Reconstructed face region expands gradually as training proceeds (odd rows in Fig. 19). The number of inlier pixels, which has a larger robust residual error than the threshold, also increases during the training (dark pixels in even row images in Fig. 19).

M Limitations of our method

Our approach is unsupervised and only relies on photometric consistency between the input image and the model, hence it is prone to fail in extreme cases (Fig. 20). The robust residual loss has an effect to expand face region so that the model can explain as many pixels as possible. Thus, skin-colour like hair causes over-expansion of a face (top-left), and skin-colour like background (top-right) causes misalignment of a face. In addition, we model the appearance only based on diffuse reflection formulated with the inverse spherical harmonic lighting. Thus, the quality of reconstruction is degraded if the input image has strong occlusion (middle-left) or strong shadow (middle-right). As our approach can reconstruct only a face, which a 3DMM can explain, extreme expressions cannot be reconstructed (bottom-left). Extreme pose is also difficult to reconstruct due to strong self-occlusion.

N Comparison between different network architectures

Fig. 18 shows comparison of outputs from different size of networks. In the main paper, we employ regular U-Net, which contains of four down-sampling

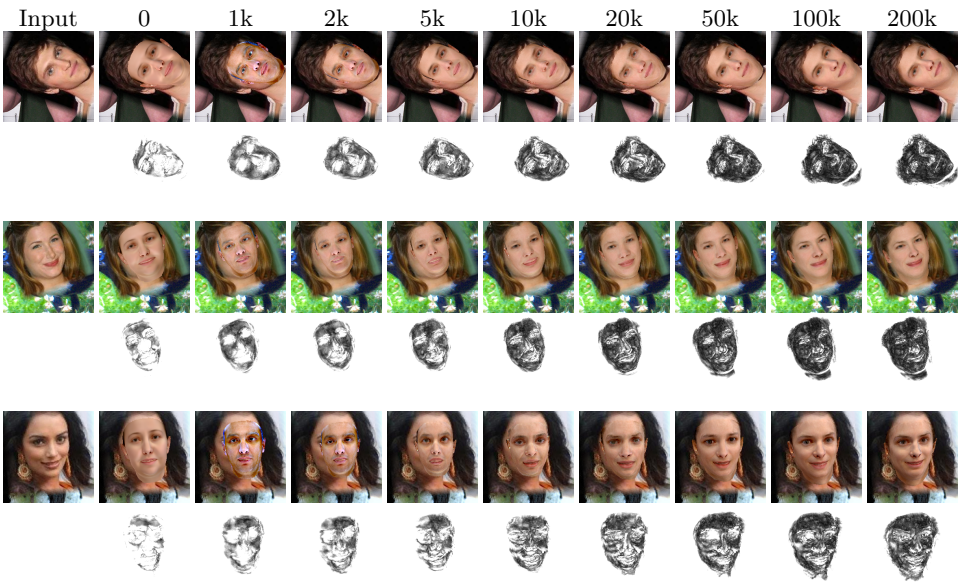


Fig. 19: Convergence of reconstructed images during training. Odd rows show the overlay of the reconstructed image. Even rows show the visualisation of the robust residual loss on each pixel. .

layers and four up-sampling layers. Each scale contains two convolution layers followed by batch normalisation and ReLU. The number of channels of an input tensor in each scale is 3, 64 128, 256, 512 for down-scaling, and 1024, 512, 256, 128, 64 for up-scaling. The number of channels of an output tensor in each scale is 64, 128, 256, 512, 512 for down-scaling, and 256, 128, 64, 64, 5 for up-scaling. We define narrow U-Net by halving the number of each channel except input and output, and we define wide U-Net by doubling the number of each channel except input and output. Additionally, we define deep U-Net by replacing double convolution in each scale by quad convolution. We qualitatively compare the quality of reconstructed images and geometry. Generally, increase of the number of channels improves the quality of reconstruction, whereas increase of the number of layer does not. In addition, we train FCN [10] with ResNet101 and DeepLab v3 [2] with ResNet101 instead of U-Net based on our approach. In our experiment, no successful training condition was found for FCN and DeepLab v3.

O Results from video sequences

We apply our approach to video sequence frame by frame. Fig. 22 shows reconstructed images and geometry for five selected frames from each sequence. These results exhibit the stability of our approach against face movement, perturba-

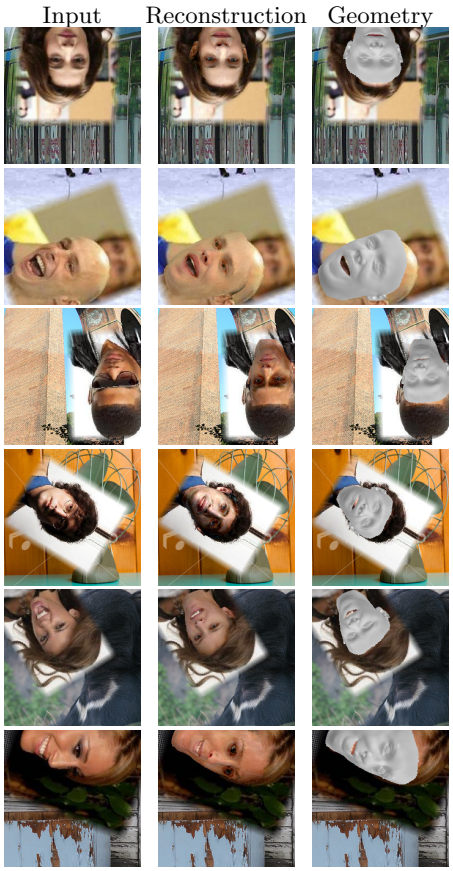


Fig. 20: Example of images which our approach fails to reconstruct a reasonable shape. Top to bottom: Skin-colour like hair, skin-colour like background, strong occlusion, strong shadow, extreme expression, and extreme pose lead to inaccurate reconstruction.

tion, and facial expression. Additionally, we provide these results in a form of video file as supplementary material. In the video material, input images, reconstructed images, reconstructed geometry, estimated correspondence, estimated confidence, and estimated depth are shown in top-left, top-middle, top-right, bottom-left, bottom-middle, and bottom-right, respectively.

P Limitations of MoFA

One contribution of our approach is to enable 3D face reconstruction from an arbitrary in-plane pose face image, which can be simulated by 2D similarity transformation of the input image. To validate our contribution, we tested MoFA

with 2D similarity data augmentation based on our MoFA [15] reimplementation (2D-augmented MoFA). Since our approach employs pretraining based on roughly aligned images, we also applied pretraining to 2D-augmented MoFA. For pretraining, we use the same images as our network. To make the training stable, we train only camera rotation and xy-translation estimation in first 10,000 iterations. After that, we train camera rotation and full translation estimation for 38,000 iterations. During pretraining, we fix 3DMM and lighting parameters, and only use landmark loss. In our experiment, we employ L1-norm instead of L2-norm as landmark loss for stability. Fig. 23 shows examples of reconstruction by the pretrained network.

Based on pretrained 2D-augmented MoFA, we investigated parameters which enable unsupervised training without landmarks. We employ statistical regularisation of 3DMM parameters and photometric loss between a reconstructed image and an input image for training. However, no successful parameter was found. We also tested training with landmark loss. Fig. 24 and Fig. 25 compare the training loss of 3 types of augmentation (no augmentation, 2D shift transformation, and 2D similarity transformation). This experiment indicates that reconstruction of a face with arbitrary in-plane pose is much more difficult than reconstruction of an aligned face for conventional CNN such as VGG19 [13] even if landmarks are provided. Therefore, it is not surprising that fragile unsupervised training fails.

Q Derivation of camera matrix

In the differentiable linear least square layer, we compute an inverse perspective camera matrix $\mathbf{P}$ and $\mathbf{q}$ such that:

$$\begin{bmatrix} \hat{\mathbf{p}}_1^t \\ \hat{\mathbf{p}}_2^t \\ \hat{\mathbf{p}}_3^t \end{bmatrix} = \mathbf{P}^{-1} = \psi \mathbf{K}\mathbf{R} \tag{19}$$

$$\hat{\mathbf{q}} = -\mathbf{P}^{-1}\mathbf{q} = \psi \mathbf{K}\mathbf{t} \tag{20}$$

where $\mathbf{K}[\mathbf{R}\,\mathbf{t}]$ represents a classical projective camera matrix. To obtain a camera matrix, we decompose $\mathbf{P}$, $\mathbf{q}$ into $\mathbf{K}, \mathbf{R}, \mathbf{t}$ as:

$$s = \|\mathbf{q}_3\|_2^2 \tag{21}$$

$$\mathbf{r}_3 = \frac{\mathbf{q}_3}{s} \tag{22}$$

$$k_5 = \hat{\mathbf{p}}_3^t \mathbf{r}_3 \tag{23}$$

$$k_4 = \|\hat{\mathbf{p}}_2 - k_5\mathbf{r}_3\|_2^2 \tag{24}$$

$$\mathbf{r}_2 = \frac{\hat{\mathbf{p}}_2 - k_5\mathbf{r}_3}{k_4} \tag{25}$$

$$k_3 = \hat{\mathbf{p}}_1^t \mathbf{r}_3 \tag{26}$$

$$k_2 = \hat{\mathbf{p}}_1^t \mathbf{r}_2 \tag{27}$$

$$k_1 = \|\hat{\mathbf{p}}_1 - k_2\mathbf{r}_2 - k_3\mathbf{r}_3\|_2^2 \tag{28}$$

$$\mathbf{r}_1 = \frac{\hat{\mathbf{p}}_1 - k_2\mathbf{r}_2 - k_3\mathbf{r}_3}{k_1} \tag{29}$$

$$\mathbf{K} = \begin{bmatrix} k_1 & k_2 & k_3 \\ 0 & k_4 & k_5 \\ 0 & 0 & 1 \end{bmatrix} \tag{30}$$

$$\mathbf{R} = \begin{bmatrix} \hat{\mathbf{r}}_1^t \\ \hat{\mathbf{r}}_2^t \\ \hat{\mathbf{r}}_3^t \end{bmatrix} \tag{31}$$

$$\mathbf{t} = \frac{1}{\psi}\mathbf{K}^{-1}\hat{\mathbf{q}} \tag{32}$$

R Derivation of the solution of the linear least squares

In our network, we solve the linear least square problem for geometry and colour. N_p pixels are sampled for the least squares. Assuming, on jth sampled pixel, $\mathbf{x}_j = (x_j, y_j, 1)^T$ is pixel coordinates, d_j is depth value, ω_j is confidence value, $\mathbf{B}_j$ is 3DMM shape basis matrix, and $\mathbf{a}_j$ is 3DMM shape mean, each element of inverse camera matrix $\mathbf{P}$, $\mathbf{q}$ and 3DMM shape coefficients $\mathbf{m}$ are derived as:

$$\mathbf{\Psi}_j = \begin{bmatrix} d_j\mathbf{x}_j^T & 1 & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & d_j\mathbf{x}_j^T & 1 & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & d_j\mathbf{x}_j^T & 1 \end{bmatrix} \tag{33}$$

$$\mathbf{\Theta} = \begin{bmatrix} \mathbf{\Psi}_1 & -\mathbf{B}_1 \\ \mathbf{\Psi}_2 & -\mathbf{B}_2 \\ \vdots & \vdots \\ \mathbf{\Psi}_{N_p} & -\mathbf{B}_{N_p} \\ \mathbf{E}_{12\times12} & \mathbf{0}_{12\times(N_s+N_e)} \\ \mathbf{0}_{(N_s+N_e)\times12} & \mathbf{E}_{(N_s+N_e)\times(N_s+N_e)} \end{bmatrix} \tag{34}$$

$$\mathbf{\Omega} = diag \begin{pmatrix} \omega_1 \\ \omega_1 \\ \omega_1 \\ \omega_2 \\ \omega_2 \\ \omega_2 \\ \vdots \\ \omega_{N_p} \\ \omega_{N_p} \\ \omega_{N_p} \\ \mathbf{0}_{3\times 1} \\ 1 \\ \mathbf{0}_{3\times 1} \\ 1 \\ \mathbf{0}_{3\times 1} \\ 1 \\ \alpha_1^2 \\ \alpha_2^2 \\ \vdots \\ \alpha_{N_s+N_e}^2 \end{pmatrix} \tag{35}$$

$$\mathbf{\Upsilon} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_{N_p} \\ \mathbf{0}_{11\times 1} \\ z_0 \\ \mathbf{0}_{(N_s+N_e)\times 1} \end{bmatrix} \tag{36}$$

$$\begin{bmatrix} P_{1,1} \\ P_{1,2} \\ P_{1,3} \\ q_1 \\ P_{2,1} \\ P_{2,2} \\ P_{2,3} \\ q_2 \\ P_{3,1} \\ P_{3,2} \\ P_{3,3} \\ q_3 \\ \mathbf{m} \end{bmatrix} = (\mathbf{\Theta}^T \mathbf{\Omega} \mathbf{\Theta})^{-1} \mathbf{\Theta}^T \mathbf{\Omega} \mathbf{\Upsilon} \tag{37}$$

where α_i represents the weight for regularisation. $\mathbf{0}$ represents zero matrix and $\mathbf{E}$ represents identity matrix.

Assuming, on jth sampled pixel, $\mathbf{i}_j$ is pixel value, F_j is inverse lighting spherical harmonic basis, ω_j is confidence value, $\mathbf{D}_j$ is 3DMM albedo basis matrix, and $\mathbf{c}_j$ is 3DMM albedo mean, spherical harmonic coefficients $\mathbf{l}$ and 3DMM colour coefficients $\mathbf{h}$ are derived as:

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{i}_1\mathbf{F}_1 & -\mathbf{D}_1 \\ \mathbf{i}_2\mathbf{F}_2 & -\mathbf{D}_2 \\ \vdots & \vdots \\ \mathbf{i}_{N_p}\mathbf{F}_{N_p} & -\mathbf{D}_{N_p} \\ \mathbf{E}_{27\times 27} & \mathbf{0}_{27\times N_r} \\ \mathbf{0}_{N_r\times 27} & \mathbf{E}_{N_r\times N_r} \end{bmatrix} \tag{38}$$

$$\mathbf{\Pi} = diag \begin{pmatrix} \omega_1 \\ \omega_1 \\ \omega_1 \\ \omega_2 \\ \omega_2 \\ \omega_2 \\ \vdots \\ \omega_{N_p} \\ \omega_{N_p} \\ \omega_{N_p} \\ \xi \\ \xi \\ \vdots \\ \xi \\ \gamma_1^2 \\ \gamma_2^2 \\ \vdots \\ \gamma_{N_r}^2 \end{pmatrix} \tag{39}$$

$$\mathbf{\Xi} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_{N_p} \\ \mathbf{0}_{27\times 1} \\ \mathbf{0}_{N_r\times 1} \end{bmatrix} \tag{40}$$

$$\begin{bmatrix} \mathbf{l} \\ \mathbf{h} \end{bmatrix} = (\mathbf{\Lambda}^T\mathbf{\Pi}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T\mathbf{\Pi}\mathbf{\Xi} \tag{41}$$

where γ_i and ξ represent the weight for regularisation.

References

1. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: Proceedings of the IEEE international conference on computer vision. pp. 1513–1520 (2013)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
3. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: IEEE Computer Vision and Pattern Recognition Workshops (2019)
4. Floater, M.S.: Parametrization and smooth approximation of surface triangulations. Computer aided geometric design **14**(3), 231–250 (1997)
5. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8377–8386 (2018)
6. Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., Vetter, T.: Morphable face models-an open framework. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 75–82. IEEE (2018)
7. Jakab, T., Gupta, A., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks through conditional image generation. In: Advances in neural information processing systems. pp. 4016–4027 (2018)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012), http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
9. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
11. Martin Koestinger, Paul Wohlhart, P.M.R., Bischof, H.: Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In: Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (2011)
12. Shu, Z., Sahasrabudhe, M., Alp Guler, R., Samaras, D., Paragios, N., Kokkinos, I.: Deforming autoencoders: Unsupervised disentangling of shape and appearance. In: Proceedings of the European conference on computer vision (ECCV). pp. 650–665 (2018)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3476–3483 (2013)
15. Tewari, A., Zollöfer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Christian, T.: MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In: The IEEE International Conference on Computer Vision (ICCV) (2017)

16. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks by factorized spatial embeddings. In: Proceedings of the IEEE international conference on computer vision. pp. 5916–5925 (2017)
17. Tran, A.T., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3D morphable models with a very deep neural network. In: Proc. CVPR. pp. 5163–5172 (2017)
18. Wiles, O., Koepke, A., Zisserman, A.: Self-supervised learning of a facial attribute embedding from video. arXiv preprint arXiv:1808.06882 (2018)
19. Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A.: Robust facial landmark detection via recurrent attentive-refinement networks. In: European conference on computer vision. pp. 57–72. Springer (2016)
20. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: European conference on computer vision. pp. 1–16. Springer (2014)
21. Zhang, Y., Guo, Y., Jin, Y., Luo, Y., He, Z., Lee, H.: Unsupervised discovery of object landmarks as structural representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2694–2703 (2018)
22. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European conference on computer vision. pp. 94–108. Springer (2014)
23. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE transactions on pattern analysis and machine intelligence **38**(5), 918–930 (2015)

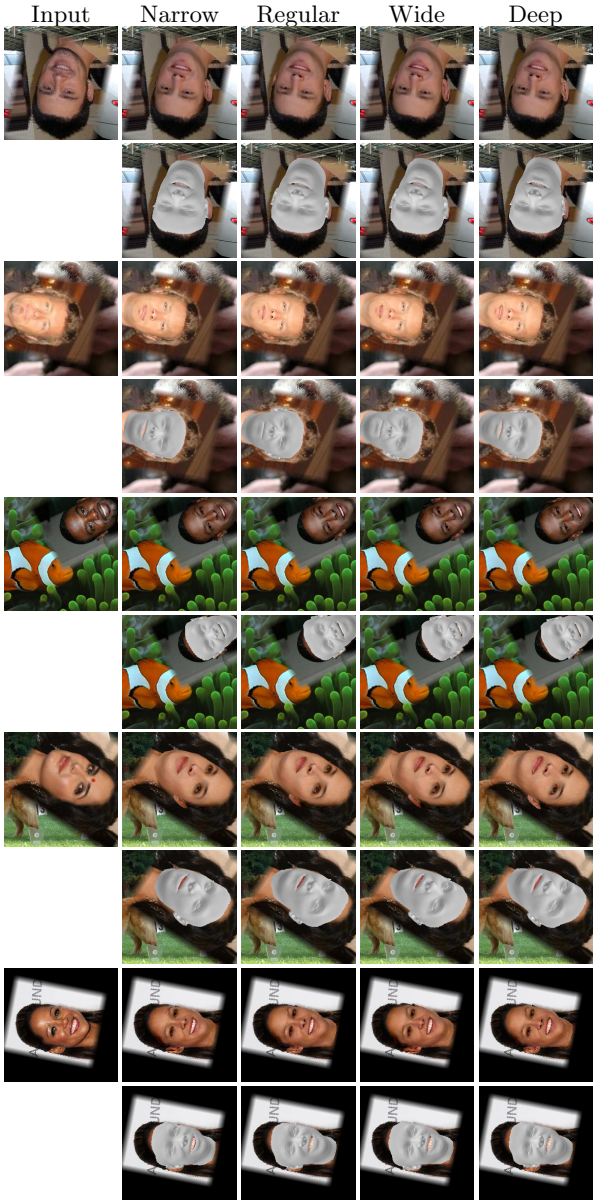


Fig. 21: Comparison of reconstruction results using different size of network. Odd rows shows reconstructed images, and even rows shows reconstructed geometry. The first column shows input images, the second column shows results of half breadth U-Net, the third column shows results of regular U-Net, the forth column shows results of double breadth U-Net, and the fifth column shows double depth U-Net.

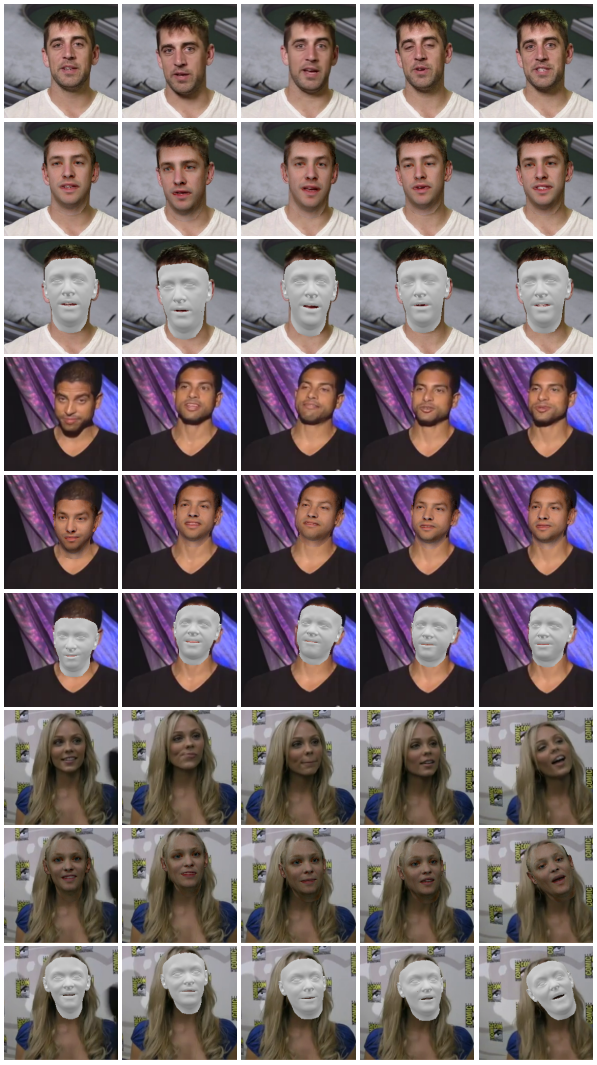


Fig. 22: Reconstruction results from video sequences. Every first rows show input images, every second row shows reconstructed images, and every third rows show reconstructed geometry.

Fig. 23: Reconstructed images by 2D-augmented MoFA [15] after pretraining.

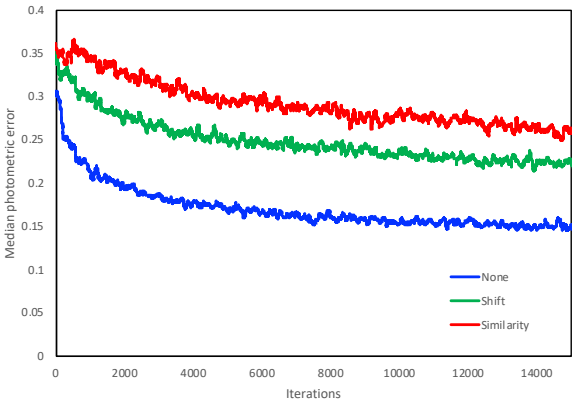


Fig. 24: Photometric loss of 2D-augmented MoFA [15] with landmark loss during training. The error is calculated as median value of training loss for each 100 iteration.

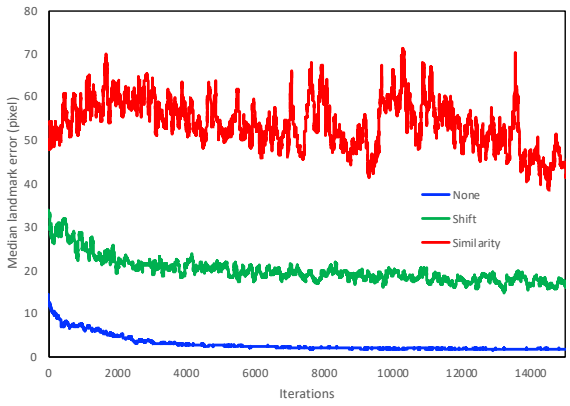


Fig. 25: Landmark loss of 2D-augmented MoFA [15] with landmark loss during training. The error is calculated as median value of training loss for each 100 iteration.