3D Face Reconstruction with Dense Landmarks

Supplementary material

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, Tom Cashman, and Julien Valentin

Microsoft

1 Additional qualitative results

Please see Figure 3 for dense landmark predictions on the entire 300W Challenging dataset, using a ResNet 101 [4].

Please see Figure 4 for qualitative comparisons between our approach and several recent methods for 3D face reconstruction in the wild.

2 $E_{\text{intersect}}$

Since our 3D face model contains separate parts for the teeth and eyeballs, intersections can occur. Though they are uncommon, we encourage the optimizer to avoid face mesh self-intersections by penalizing skin vertices that enter the convex hulls of the eyeballs or teeth parts.

$$E_{\text{intersect}} = E_{\text{eyeballs}} + E_{\text{teeth}}$$

We attach a sphere of fixed radius to each eyeball center. For each eyelid skin vertex that falls inside its corresponding eyeball sphere, E_{eyeballs} penalizes the squared distance between that vertex to the sphere's exterior surface. Since this is trivial to implement, we will omit the details.

However, the teeth cannot be well-represented with a simple primitive like a sphere, so E_{teeth} is more complicated. Instead, we represent the upper and lower teeth parts each with a convex hull of J planes defined by normal vector $\hat{\mathbf{n}}_j$ and distance to origin p_j. Lets say I represents a set of lip vertices we wish to keep outside one of these convex hulls.

$$E_{\text{teeth}} = \sum_{i \in I} D_i^2$$

Where D_i measures the distance the i^{th} skin vertex is inside the convex hull,

$$D_i = \min_{j \in J} \{d_{i,1}, \ldots, d_{i,J}\}$$

and $d_{i,j}$ measures the internal distance between the i^{th} skin vertex and the jth plane of the convex hull.

$$d_{i,j} = -\min\left(\hat{\mathbf{n}}_j \cdot \mathbf{x}_i + p_j, 0\right)$$

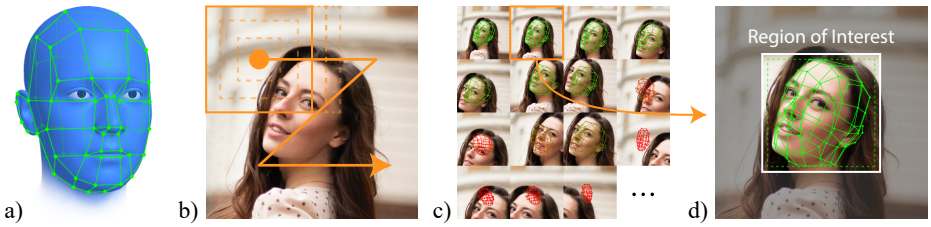

Fig. 1. To extract a distortion-free Region-of-Interest (ROI) of the head we run a full-head sparse landmark CNN (a) on multi-scale sliding windows across the full image (b), take the landmarks from the most confident window (c), inscribe an expanded square around them, and use it as our ROI (d) for our dense landmark CNN.

Fig. 2. When running in real time, we extract ROIs using an affine transform. The eyes and mouth, shown as green points in (a) are remapped every frame to a fixed triangle in ROI space (b). The resulting rotated ROI rectangle, shown in red in (a) is remapped into the square (b).

3 Region-of-Interest extraction

Our method is "top-down", so it relies on being provided a reasonable Region-of-Interest (ROI) of a face. During offline processing, where compute is not paramount, we prefer to extract distortion-free ROIs from a full image, with zero rotation and uniform scaling (see Figure 1).

When running in real time, we modify our ROI strategy to perform more data normalization, producing ROIs where the eyes always appear in the same place, and the mouth always appears on the same horizontal line (see Figure 2). This corresponds to a rotation and non-uniform scaling. While this makes the task easier for the dense landmark CNN (preferable for low-capacity neural networks), it breaks down for profile faces, so we prefer the distortion free approach described in Figure 1 when compute is not constrained.

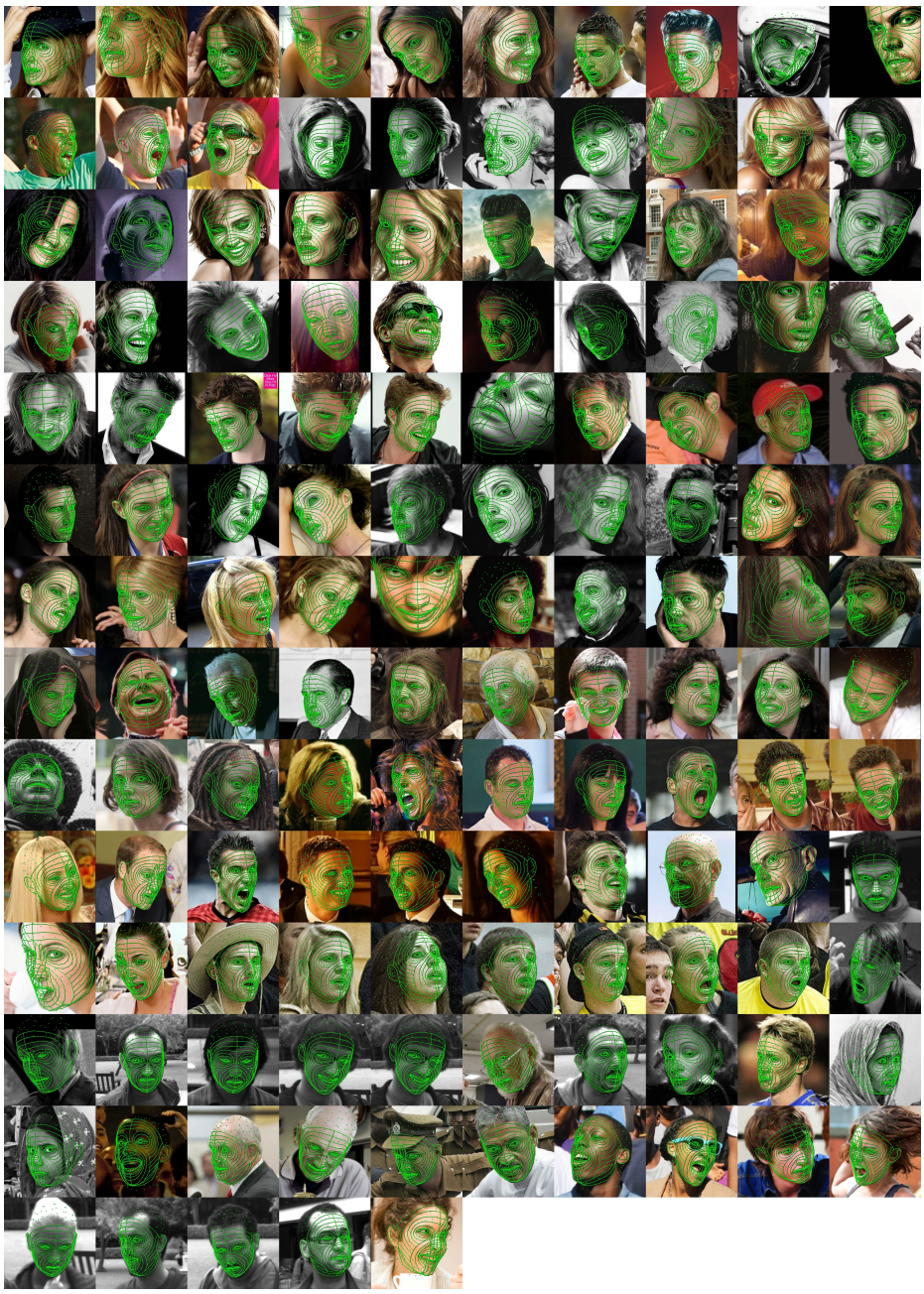

Fig. 3. Dense landmark predictions for all of the images in 300W Challenging dataset. Note the accuracy of our landmark model in challenging scenarios including extreme expressions, occlusion, pose variation, lighting variation, and poor image quality.

Fig. 4. Further qualitative comparisons between our approach and publicly available recent previous methods for monocular 3D face reconstruction.

4 Additional implementation details

Training details. Both (a) the 703 landmarks model with pretrained ResNet 101 [4] backbone for offline multi-view fitting and (b) the 320 landmarks model with pretrained MobileNet V2 [7] backbone for real-time monocular fitting, were trained with batch size 128 and learning rate schedule StepLR(step_size=100, gamma=0.5). We used AdamW [5] in the PyTorch library [6] with default optimizer hyperparameters.

Focal length initialization and optimization. If the focal length of the camera is known, we keep it fixed during 3D model fitting. Otherwise, if it is unknown, we initialize it to be 45 degree effective horizontal FOV, and optimize it together with other 3D face model parameters. For simplicity, we assume the principal point is at image center and focal length value (in pixels) is the same for x and y direction (square pixels) Empirically, we found that these assumptions were acceptable.

References

1. 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: CVPR Workshops (2019)
2. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. ACM Transactions on Graphics (ToG) (2021)
3. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards Fast, Accurate and Stable 3D Dense Face Alignment. In: ECCV (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
5. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)
6. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
7. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobilenetV2: Inverted residuals and linear bottlenecks. In: CVPR (2018)
8. Sanyal, S., Bolkart, T., Feng, H., Black, M.: Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In: CVPR (2019)
9. Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In: ECCV (2020)