Continuous Landmark Detection with 3D Queries - Supplemental Material

Prashanth Chandran Gaspard Zoss Paulo Gotardo[*] Derek Bradley
DisneyResearch|Studios
prashanth.chandran,gaspard.zoss,derek.bradley@disneyresearch.com, gotardop@gmail.com

1. Architecture and Training details

As described in the main paper, the modular nature of our network enables us to easily replace or resuse various building blocks for the Feature extraction network $\mathcal{F}$, and the queried landmark predictor $\mathcal{P}$ to construct multiple variants of our network. We particularly evaluted 3 combinations of $\mathcal{F}$ and $\mathcal{P}$ in our work: i) ConvNext + MLP ii) ConvNext + Transformer, and iii) MobileNetV3 + MLP. In this section, we describe the architectural details of each of our variants.

1.1. Feature Extraction Networks

We use the official implementation of the ConvNext encoder [4] and select the *small* configuration. This configuration corresponds to a ConvNext encoder with depths: [3, 3, 27, 3] and features: [96, 192, 384, 768]. The output of the network is a 768 dimensional descriptor of the input image, obtained by average pooling the features at the last convolutional layer of the encoder. We make no modifications to the ConvNext Encoder. For our real time variant, we use the MobileNetV3 backend as the feature extraction network. We use the implementation of MobileNetV3 made available as part of the pyTorch library [7], *mobilenet_v3_large*. We do not use pre-trained weights, and train from scratch to specialize only on face images. The ConvNext encoder has a capacity of roughly 50 million parameters and MobileNetV3 consists of about 6 million parameters.

1.2. Queried Landmark Predictors

We evaluate two versions of the queried landmark predictor $\mathcal{P}$. The first variant of $\mathcal{P}$ is a simple 4 layer MLP consisting of linear layers and *GeLU* activations. We show a detailed breakdown of this MLP in Fig. 1 (left). The second variant of our query prediction network consists of Transformer blocks. While the MLP can process a batch of multiple queries at once, independently, the transformer uses self attention to exchange information across the simultaneous queries and leverage correlations across output landmark positions. This version of $\mathcal{P}$ is seen in Fig. 1 (right). In this transformer variant, the number of simultaneously queried positions on the canonical shape corresponds to the sequence length of the transformer, for modeling *spatial* correlations. Therefore the input to this transformer queried landmark predictor is a sequence on tokens, where each token is a concatenation of a query q_j and the image feature descriptor f_i. Note that the image descriptor is replicated n times and concatenated with each query $[q_0, q_1, ..q_n]$. As the memory required to build a full self-attention matrix across the input queries grows quadratically with the sequence length n, researchers have proposed multiple alternative designs that reduce the memory footprint, in which self-attention increases only linearly with sequence length [2, 9]. In our work, we use the *XCiT* attention block [2] with an intermediate feature dimension of 768. In both the MLP and transformer variants of our predictor $\mathcal{P}$, the learned position encoder $\mathcal{M}$ has the same design, consisting of a two-layer MLP with *GeLU* activations.

1.3. Optimization

We train our networks ($\mathcal{F}$, $\mathcal{M}$, and $\mathcal{P}$) end to end using the AdamW optimizer [5] with learning rate 1e-4, and other hyper-parameters set to default. We trained our models on a single NVidia A6000 GPU for 3 days, with batch size 64.

2. Additional Evaluations and Results

We now present additional results and validations, including an evaluation of occlusion contour landmarks, an ablation of different canonical face shapes, a comparison of our architecture variants, additional anatomy results, additional 3D reconstruction results, and an experiment to show how any dense supervision can be used to improve accuracy with our method.

2.1. 300-W Evaluation and Occlusion Contours

In our method, the 3D points on the canonical model do not necessarily map one-to-one to 3D points on the face, which would cause occluded points over the cheek in half profile. Instead, our model learns to map 3D points on the canonical to "semantic" 2D pixels, which could be either fixed points on the face or sliding points for occlusion contours - *entirely depending on what data we train on*. We il-

[*] Now at Google

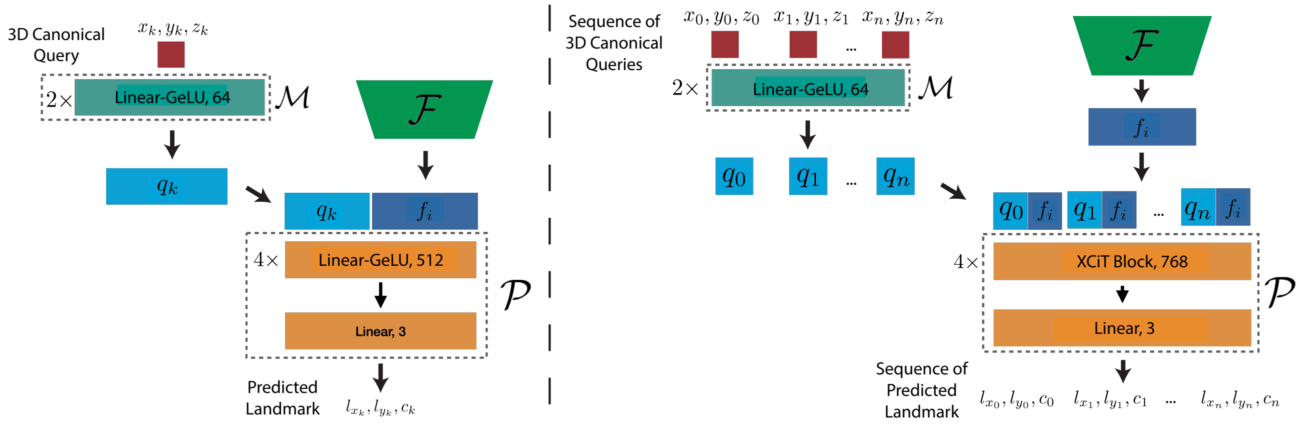

Figure 1. Architecture breakdowns of the two variants of our queried landmark predictor $\mathcal{P}$. Left: An MLP variant. Right: A Transformer variant consists of 4 blocks of *XCiT* self-attention.

lustrate this in Fig. 2, by showing the result of training only on 300-W [8], which contains sliding occlusion contours, versus training only on fake-it-till-you-make-it [10], which contains fixed 3D landmarks (not sliding). Our method will faithfully reproduce the behavior of the training data.

2.2. Canonical Shape

We train all our main results with one canonical shape (shown in Fig. 3, left). This shape is a template face with open mouth, eyeballs, skull, jaw and teeth. While any shape can be used as the canonical, we evaluate an alternate canonical shape with closed mouth expression (see Fig. 3, right), but found that it can lead to an underestimation of the lower lip landmarks on images with large open mouth expressions, presumably caused by the change in topology as encountered by Park et al [6].

2.3. Effect of Different Architectures

In Fig. 4, we show a qualitative comparison between our 3 architectural variants on a test video, when predicting dense (500) facial landmarks. The supplemental video also provides a qualitative comparison, with different landmark layouts, where temporal stability can be appreciated. Referring to the finding of Table 2 and Fig 14 in the main paper, we find that the ConvNext + Transformer variant achieves the highest accuracy, while our MobileNetV3 + MLP variant is the fastest at inference time.

2.4. Additional Anatomy Results

We include qualitative examples of teeth landmark predictions in Fig. 5. Our method is able to predict plausible teeth positions even when the teeth are occluded, for instance, during speech. The predicted 2D teeth positions are temporally stable enough to place a template 3D model as seen in the second row of Fig. 5. The reader is kindly referred to supplementary video for several examples of estimating plausible 3D anatomy using the 2D anatomical landmarks predicted by our network. To track 3D anatomy on a video given only our 2D landmark estimates, we optimize for an isotropic scale (estimated only on a reference frame) and per-frame rigid transformations on template skull, jaw, eyes and teeth meshes while using the estimated 2D landmarks as re-projection constraints. For our anatomy fitting results, we do not use a parametric shape model and also do not enforce any temporal smoothness during the optimization. We note however that any such additional complexity can be trivially added to our method given it's ability to predict to dense landmarks, and will naturally improve the quality of the estimated anatomy.

2.5. 3D Reconstruction

In Fig. 6, we show additional qualtiative results for monocular in-the-wild 3D reconstruction using the dense landmarks predicted by our method. We use the anatomical local face model by Wu *et al.* [11] as an underlying parametric face model and fit this model to 10,000 2D landmarks predicted by our network on several in-the-wild videos. Any other parametric model, such as the FLAME [3] can be trivially used along with our method as well. We highlight the robustness of our 2D landmark predictor to challenging expressions, harsh lighting, and rapidly varying backgrounds, thereby allowing the parametric shape model to recover-high quality geometry from unconstrained video.

2.6. Dense Landmark Detection Using FaceScape

Because our approach treats landmarks as continuous queries, it enables our network to already smoothly interpolate between supervised queries on in-the-wild videos even when trained with only sparse supervision (see rows 1 and

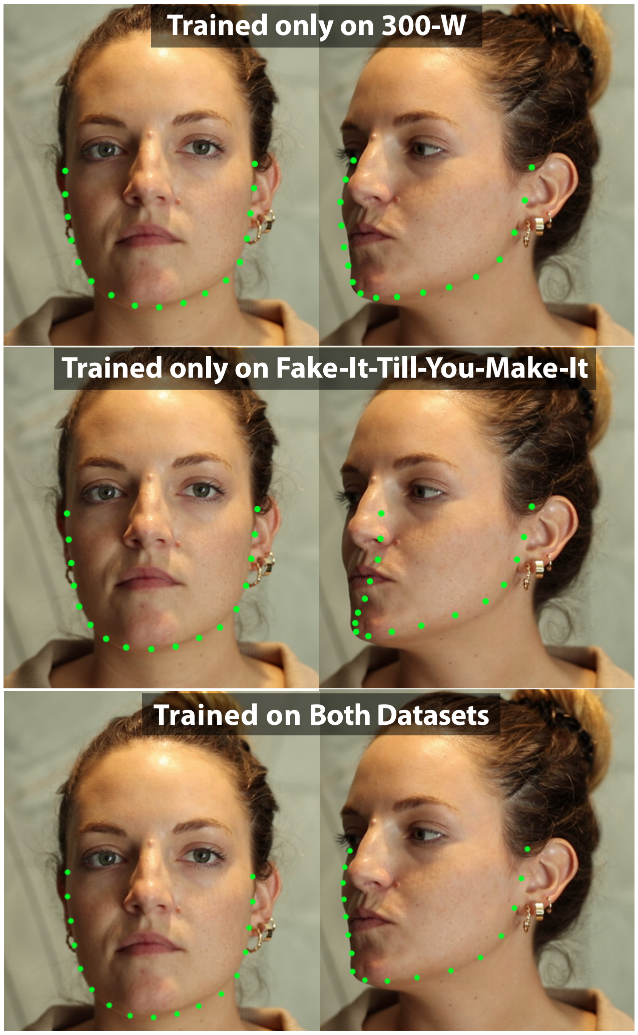

Figure 2. Our method trained only on 300-W (left) vs. only on Fake-It (right). Our method can handle both fixed and sliding landmarks as we learn to capture the trend seen in the training data, if it includes sliding. Here, on the top row we show the landmarks predicted by our method when trained only using the 300-W dataset, which contains jaw landmarks that slide along the occlusion boundary of the face. In the second row, we show the result of predicting landmarks when only training on the synthetic dataset from Wood et al. [10] which contains perfect ground truth, without sliding. Our method reflects the behavior found in the training dataset in both scenarios.

2 of Fig. 7). However, in the absence of denser supervision, such a model extrapolates unseen queries sub-optimally. As mentioned before, one of the key advantages of our method is that it can be trained on multiple, simultaneous datasets with inconsistent landmark numbers and layouts. In Fig. 7, we demonstrate how adding any dense supervision from strictly studio datasets [1, 12] improves the query extrapolation on in-the-wild images.

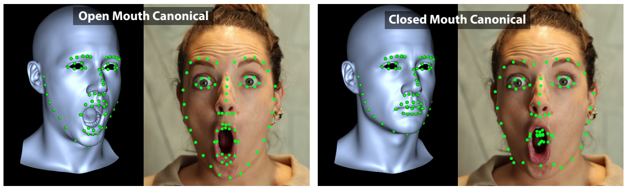

Figure 3. Effect of Changing the Canonical Shape. Empirically we found the open mouth canonical shape to give better results, especially around the mouth area.

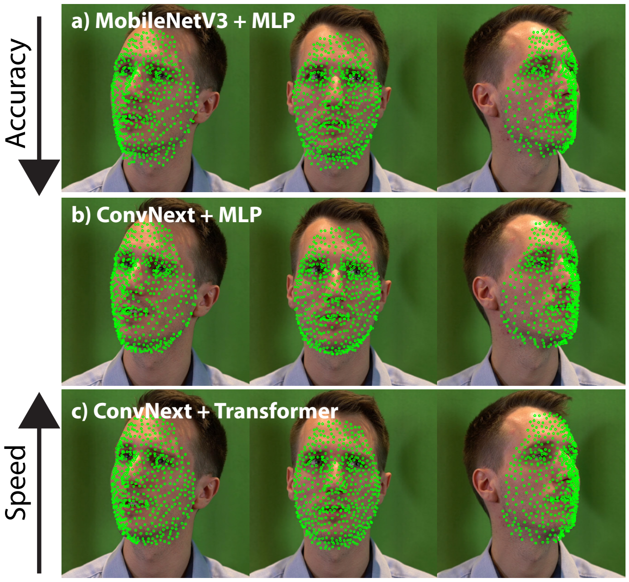

Figure 4. A qualitative comparison of our three proposed variants. Top: MobileNetV3 + MLP, Middle: ConvNext + MLP, Bottom: ConvNext + Transformer. Our ConvNext + MLP variant provides a middle ground between runtime and accuracy.

3. Acknowledgements

We wish to thank all actors who performed for us in this work.

References

[1] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Semantic deep face models. In *3DV*, pages 345–354, 2020. 3, 5

[2] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. 2021. 1

[3] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM*, 2017. 2

[4] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11966–11976, 2022. 1

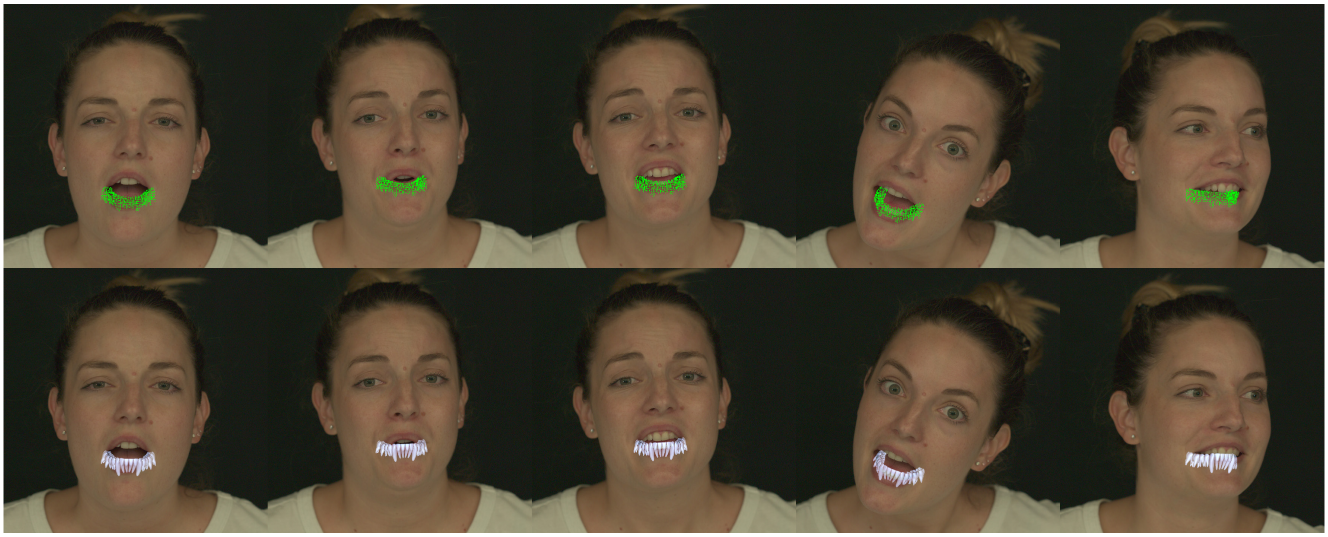

Figure 5. Top: 500 lower teeth landmarks predicted per frame as an actor performs an expressive speech. Notice the ability of our method to predict plausible teeth positions even when the teeth are not visible in the image. Bottom: the estimated 2D landmarks can be used to place a 3D template of the teeth on each frame of the performance and can provide plausible results.

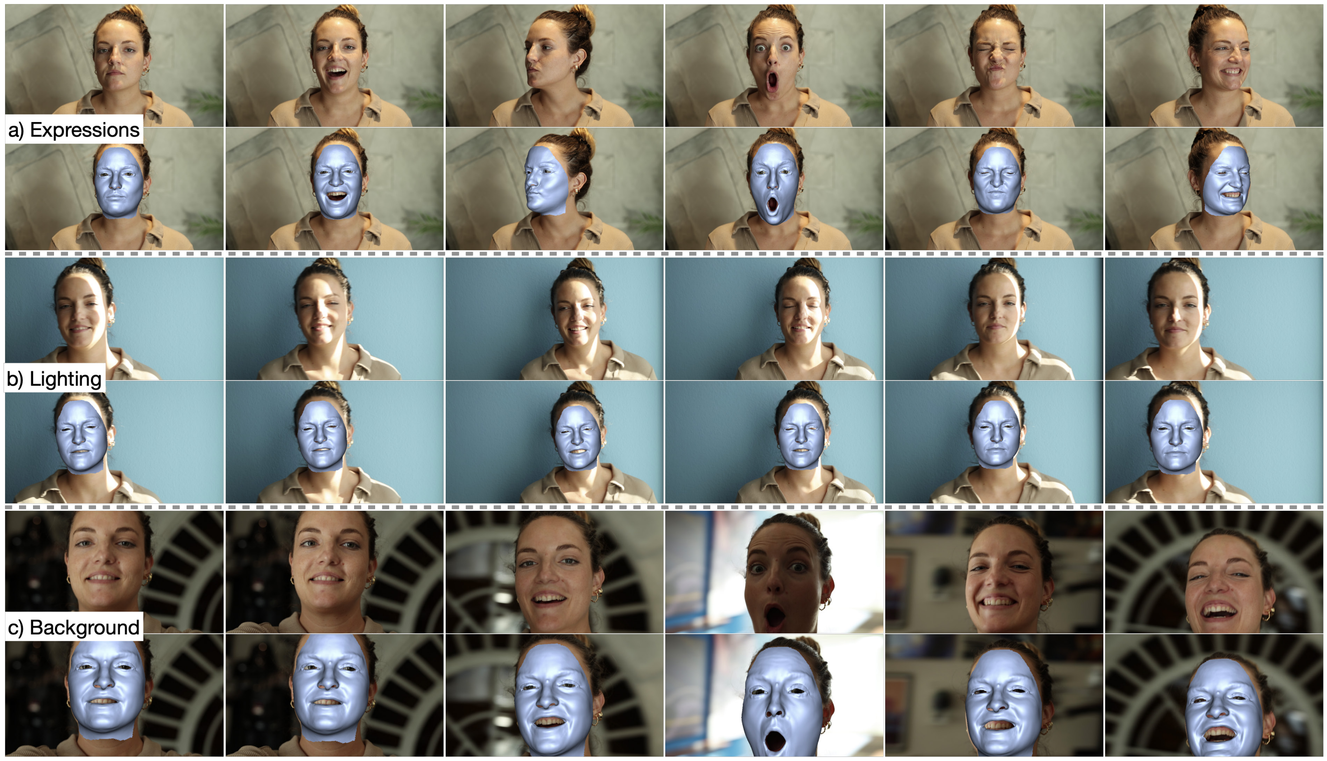

Figure 6. We show examples of in-the-wild monocular face reconstruction using 10,000 2D landmarks predicted by our model. Our network is robust to challenging expressions (a), lighting (b), and varying background, head framing, and lighting (c).

[5] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, 2017. 1

[6] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 2

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer,

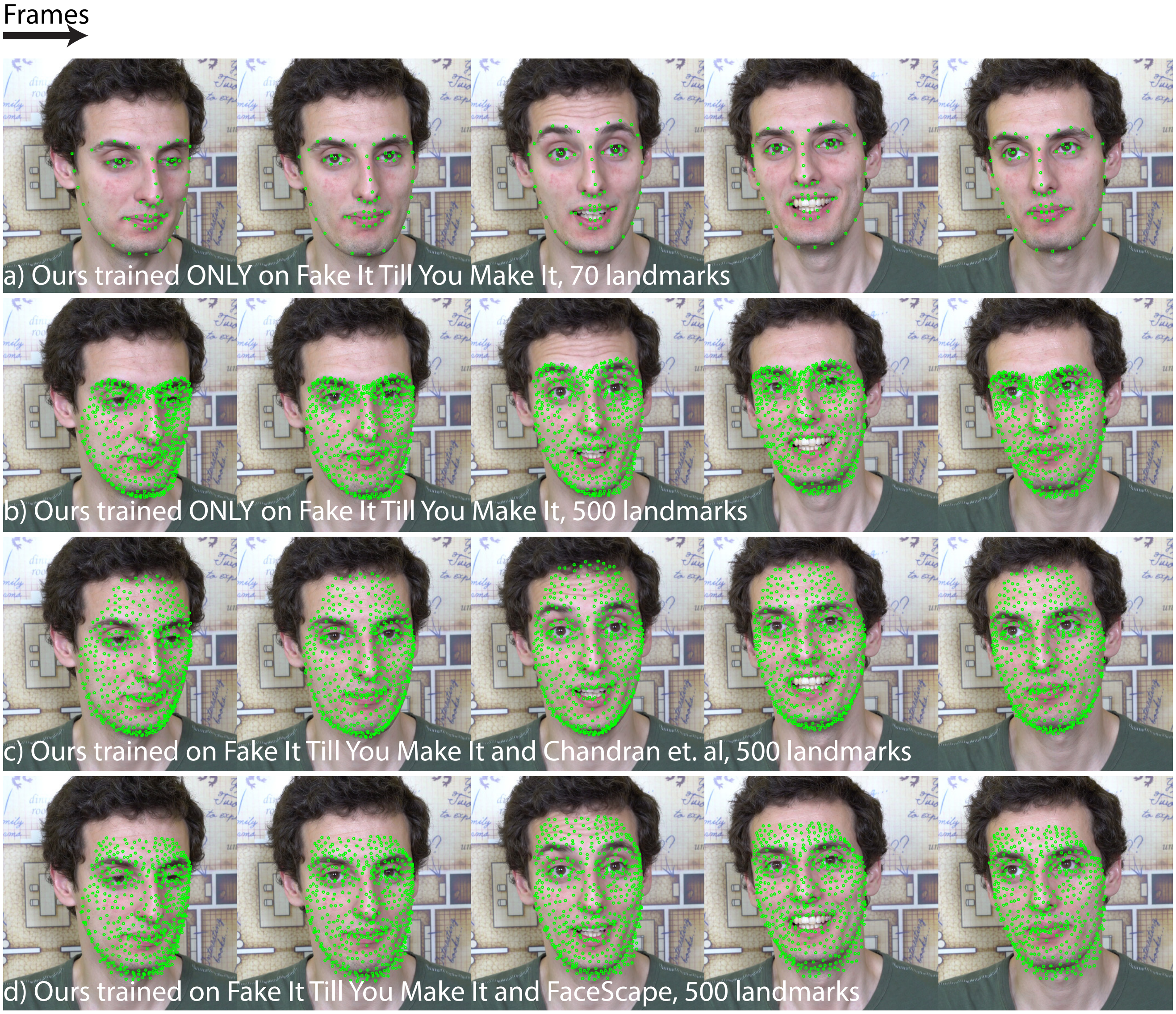
Frames
a) Ours trained ONLY on Fake It Till You Make It, 70 landmarks
b) Ours trained ONLY on Fake It Till You Make It, 500 landmarks
c) Ours trained on Fake It Till You Make It and Chandran et. al, 500 landmarks
d) Ours trained on Fake It Till You Make It and FaceScape, 500 landmarks

Figure 7. In the first row a), we show the result of training our method ONLY on the *Fake-It-Till-You-Make-It* dataset [10] and predicting the same set of 70 landmarks on a test video. The second b) shows the result of quering 500 dense landmarks from our model trained only using 70 landmarks. This leads to poor extrapolation all over, especially visible in the forehead region. In the third and fourth rows c) and d), we show qualitative results of how adding dense skin supervision from controlled studio datasets [1, 12] improves query extrapolation in in-the-wild scenarios.

James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035. 2019. 1

[8] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshops*, pages 397–403, 2013. 2

[9] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. 2020. 1

[10] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Tom Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *ICCV*, 2021. 2, 3, 5

[11] Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. An anatomically-constrained local deformation

model for monocular face capture. *ACM TOG*, 35(4), 2016. 2

[12] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *CVPR*, pages 598–607, 2020. 3, 5