Fake it till you make it: face analysis in the wild using synthetic data alone

Supplementary material

Erroll Wood* Tadas Baltrušaitis* Charlie Hewitt
Sebastian Dziadzio Thomas J. Cashman Jamie Shotton
Microsoft

1. Ablation studies

We perform a number of ablation studies, to investigate the effect of varying dataset generation parameters on downstream task accuracy.

1.1. Experiment methodology

In all experiments for model training we use a $10,000$ image synthetic dataset (split to $9,000$ training and $1,000$ validation samples). We evaluate on a landmark detection task on 300W dataset, following the label adaptation procedured described in the main paper. We use a ResNet-34 model as the backbone and train all models for 120 epochs and pick the model that performs best on the synthetic validation set.

1.2. Pose variability

In this experiment we vary the pose variability in the dataset. In full pose variability, the neck pose (in degrees) varies in the range of $(-10, 10)$ pitch $(-25, 25)$ yaw, $(-10, 10)$ roll; and head pose varies in the range of $(-30, 30)$ pitch, $(-50, 50)$ yaw, and $(-15, 15)$ roll. Further the camera is positioned with spherical coordinates, with polar angle varying between of $(-45, 45)$, and azimuthal angle varying between $(-25, 25)$. All pose values are sampled from a truncated Gaussian distribution. For this ablation study we dampen the pose variation to $0, 25\%, 50\%$, and 75% of the original values. Results can be seen in Table 1. Figure 1 shows the effect of increasing pose variation: we see more faces in profile and other non-frontal poses.

From the results it can be clearly seen that pose variation is critical for landmark detection accuracy. However, while landmark detection accuracy increases consistently for the challenging subset, it plateaus for common and private sets. This is likely due to them containing less pose variation. This demonstrates a strength of synthetic data, as it is easy to tailor the training set to contain the amount of pose variation present in test data.

*Denotes equal contribution.

Table 1. Landmark localization results on the common, challenging, and private subsets of 300W with different amount of pose variation in training data. Lower is better in all cases. Error is reported as mean of normalized mean error

Method	Common NME	Challenging NME	Private NME
Pose: 0%	3.98	9.74	5.99
Pose: 25%	3.34	6.51	4.66
Pose: 50%	**3.24**	5.14	4.25
Pose: 75%	3.24	5.07	**4.23**
Pose: 100%	3.28	**5.04**	4.28

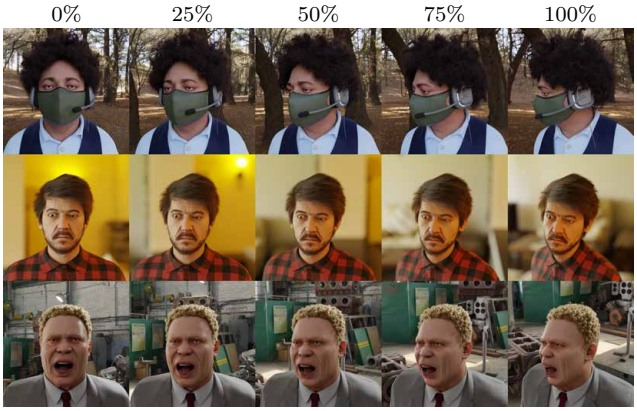

Figure 1. Examples from datasets with reduced pose variability.

1.3. Expression variability

In this experiment we vary the number of expressive faces (as opposed to neutral faces) present in the synthetic dataset. We generate datasets with 10%, 25%, 50%, 75%, and 100% neutral face images. In all experiments in the main paper, we use 10% neutral frames.

Results can be seen in Table 2. From results we see that some expression variation is really important, however, the gains in performance saturate. This is likely to the test sets not actually containing many examples of varied or extreme expressions.

Table 2. Landmark localization results on the common, challenging, and private subsets of 300W with different amount of expression variation in training data. Lower is better in all cases. Error is reported as mean of normalized mean error

Method	Common NME	Challenging NME	Private NME
Neutral: 100%	3.55	5.80	4.94
Neutral: 75%	3.30	5.05	4.32
Neutral: 50%	3.28	5.04	4.31
Neutral: 25%	**3.28**	**5.01**	**4.28**
Neutral: 10%	3.28	5.04	4.28

Table 3. Landmark localization results on the common, challenging, and private subsets of 300W with different amount of identity variation in training data. Lower is better in all cases. Error is reported as mean of normalized mean error

Method	Common NME	Challenging NME	Private NME
Identities: 25	3.43	5.18	4.48
Identities: 50	3.39	5.19	4.47
Identities: 100	3.35	5.09	4.41
Identities: 1000	3.29	5.04	4.25
Identities: 2000	3.28	**5.01**	4.31
Identities: 5000	**3.27**	5.02	**4.25**
Identities: 10000	3.28	5.04	4.28

1.4. Identity variability

In this ablation study we evaluate the effect on varying identity in training data. We keep identity (geometry, texture, accessories) fixed while varying pose, expression, camera, and environment. This allows us to see how much it is important to have identity variability compared to other types. Further, as assembling an identity is computationally expensive, we can save compute by having an identity fixed but varying other scene parameters, an example of this can be seen in Figure 2. We generate the entire 10000 image dataset, but constrain it to contain only a set number of identities.

Results of this experiment can be seen in Table 3. From the results we can see that going over 2000 unique identities provides a limited gain in performance. This could be due to limitation in variability of the underlying assets, meaning that a new unique identity does not provide sufficient additional diversity.

1.5. Render quality

As a final ablation study we explore the effect on rendered image quality. We did this by varying the number of path tracing samples used by the Cycles renderer. This leads to faster, but lower quality renders. Examples of renders with different number of Cycles samples can be see in Figure 3. For dataset used in the main paper we use 256 cycles samples with a denoiser applied on the image.

You can see the results of this experiment in Table 4. Increasing the sample count has a positive impact on landmark detection accuracy all the way up to 256 samples. This confirms that rendered image quality is important for downstream machine learning accuracy.

Figure 2. Example of keeping an identity constant, while varying expression, pose, camera, and environment

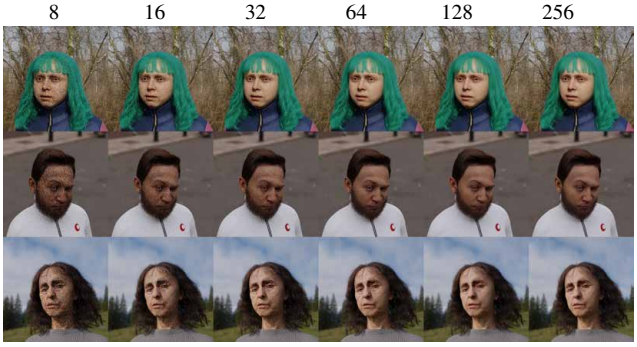

Figure 3. Examples of renders with different path tracing samples.

Table 4. Landmark localization results on the common, challenging, and private subsets of 300W with different amount of identity variation in training data. Lower is better in all cases. Error is reported as mean of normalized mean error

Method	Common NME	Challenging NME	Private NME
Samples: 8	3.56	5.45	4.65
Samples: 16	3.45	5.27	4.52
Samples: 32	3.39	5.23	4.47
Samples: 64	3.31	5.07	4.32
Samples: 128	3.29	5.03	4.33
Samples: 256	3.29	**4.99**	4.28
Samples: 256 + denoising	**3.28**	5.04	**4.28**

2. Method details

In this section we provide additional implementation details for certain stages of our face generation process.

2.1. Displacement maps

To improve realism, we apply coarse and meso-level displacement to our face geometry. Both these displacement

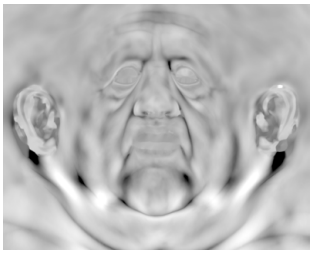


Figure 4. An example of a coarse displacement map and a meso-level displacement map for a face scan.



Figure 5. From left to right: a synthetic eye without any makeup, the same eye with eyeliner, and the same eye with eyeliner and purple metallic eye shadow.

textures are of the same resolution as the albedo texture: (8192×8192 px). Please see Figure 4 for an example of a coarse and meso-level displacement map. Note how the coarse displacement encodes the broad wrinkles of the face, while the meso-level map encodes skin-pore and fine-wrinkle geometry.

Using Blender, we first subdivide the surface of the face three times using Catmull Clarke [1] subdivision[1], to increase its vertex resolution. We then apply the coarse displacement map using a displacement modifier[2]. However, since our meso-level displacement is too fine-grained to be represented with geometry (without excessive levels of subdivision), we turn the grayscale bump-map into a normal map, and use this normal map to adjust the surface normals used during material shading.

2.2. Makeup

We randomly apply make-up effects on top of our base skin shader by layering two additional components on top: eyeliner and eye shadow. Note, we do not simulate mascara (eyelash makeup). These makeup effects have their own shaders, which are overlaid on top of the skin shader using alpha-masks which were hand-painted by an artist. We 13 eyeliner masks and 18 eye shadow masks. While eyeliner is always black, eye shadow can have a color and a degree of glitter, creating small specular reflections on the upper and lower eyelids. Please refer to Figure 5 for an example.

[1] https://docs.blender.org/manual/en/latest/modeling/modifiers/generate/subdivision_surface.html

[2] https://docs.blender.org/manual/en/latest/modeling/modifiers/deform/displace.html#displace-modifier

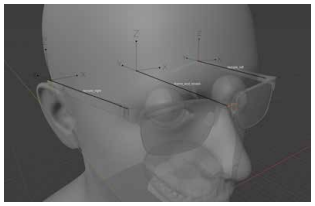


Figure 6. Eyeglasses are posed using an inverse-kinematics rig, posed using the three controllers for the nose-bridge and ears, visible on the left in orange.

2.3. Eyeglasses IK rig

Most of our face accessories represent "soft" items that are made out of some sort of fabric, e.g. t-shirts or hats made from cotton or leather. We first author these on the template face, and deform these in a non-rigid fashion using a lattice deformation field[3], driven by differences between a sampled identity, and the face template.

However, since eyeglasses are generally made of a few rigid parts, we pose these using an Inverse Kinematics (IK) rig. The position of the glasses frame is first determined using the main controller, located at the nose-bridge. The two secondary controllers are then placed just above the ears, and these determine the tilt of the glasses' temples (a.k.a. arms). Please see Figure 6.

3. Additional results

Please see the following figures for additional qualitative results for face parsing and facial landmark localization.

References

[1] E. Catmull and J. Clark. Recursively generated b-spline surfaces on arbitrary topological meshes. *Computer-aided design*, 10(6):350–355, 1978. 3

[3] https://docs.blender.org/manual/en/latest/modeling/modifiers/deform/lattice.html

Figure 7. Pairs of landmark predictions from networks trained on real (top) and synthetic (bottom) data, for the 300W dataset.

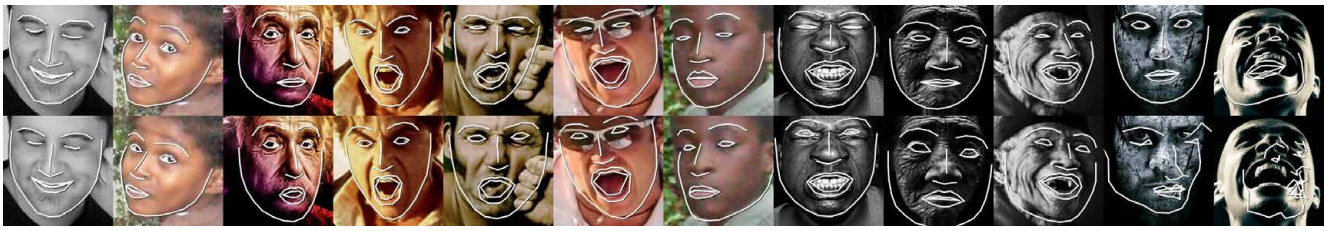

Figure 8. Pairs of landmark predictions from networks trained on real (top) and synthetic (bottom) data. Shown here are failure modes that arise when training with synthetic data only. Problematic cases include extreme illumination and uncommon skin appearance.

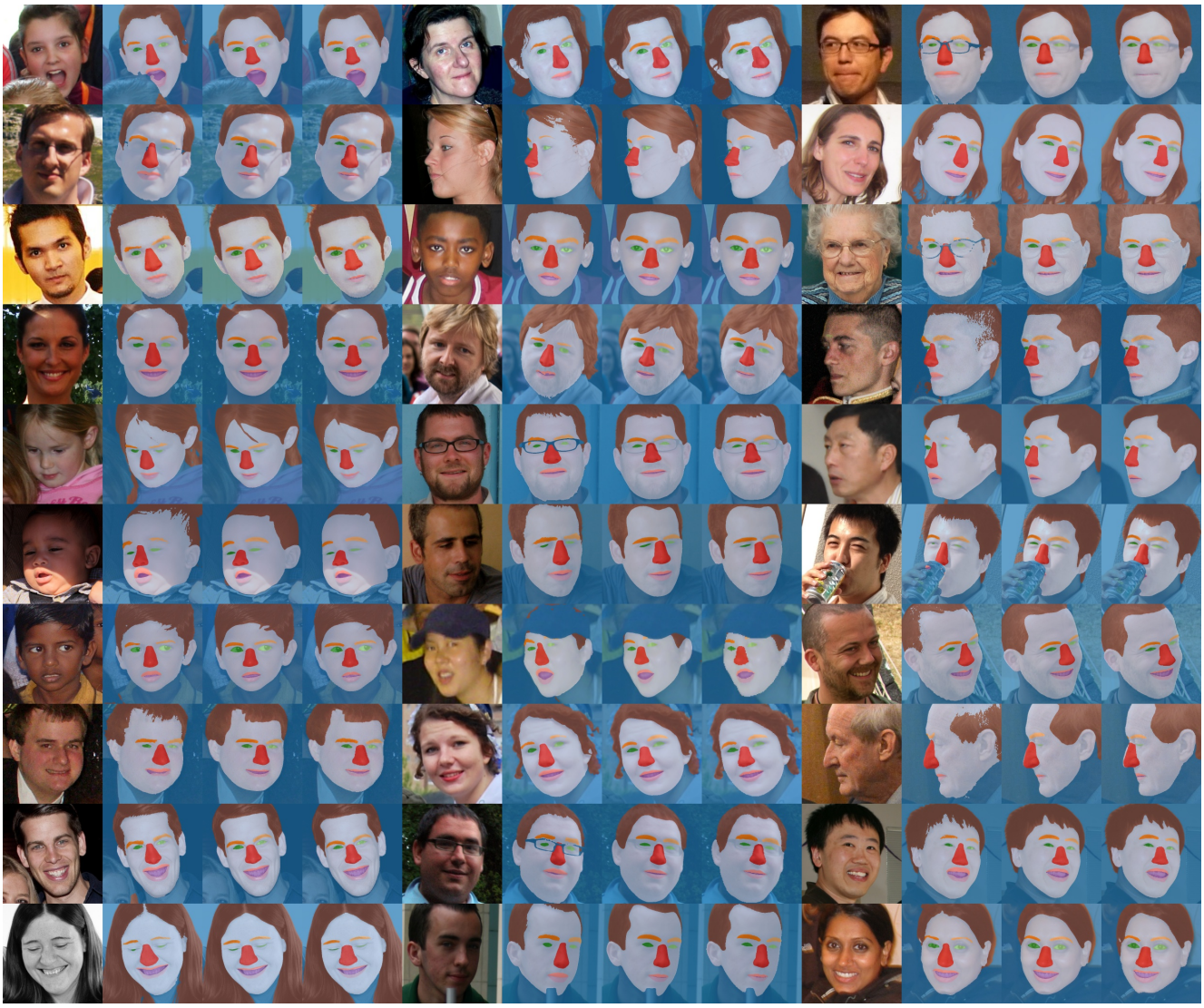

Figure 9. Shown here are face parsing results on the LaPa dataset. For each input image, shown left to right are: input color image, face parsing result from training with synthetic data only, face parsing result after label adaptation, and ground truth.

Figure 10. Additional face parsing results on the LaPa dataset, highlighting failure modes from training with synthetic data. For each set of images, shown left to right are: input color image, face parsing result from training with synthetic data only, face parsing result after label adaptation, and ground truth. Failures occur for challenging examples including unusual headwear, extreme make-up, and statue faces.

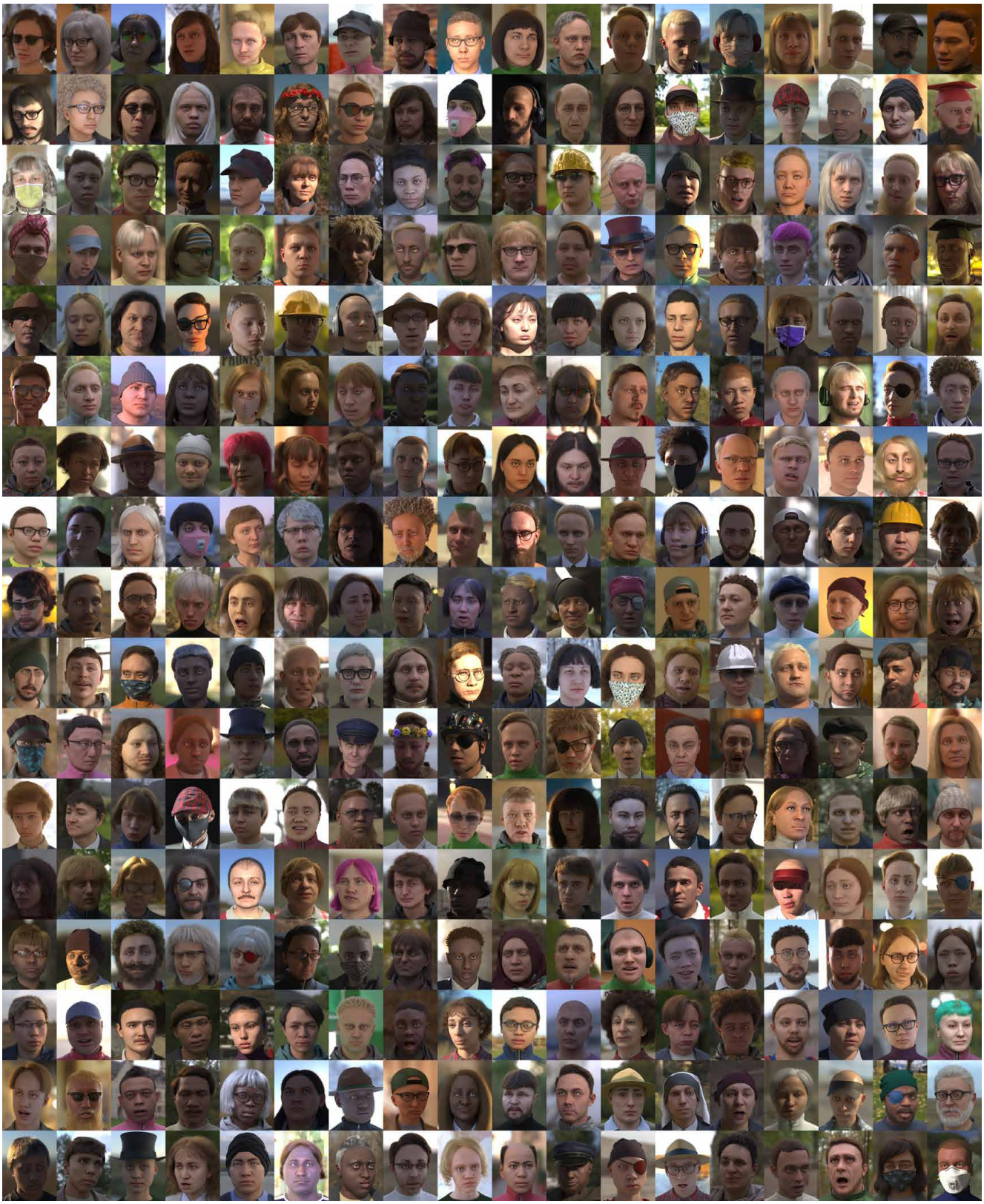

Figure 11. Here are some additional sampled synthetic faces that we have procedurally generated. Please note the wide range of appearance diversity arising from randomly combining different face shapes, textures, hair styles, and clothing.

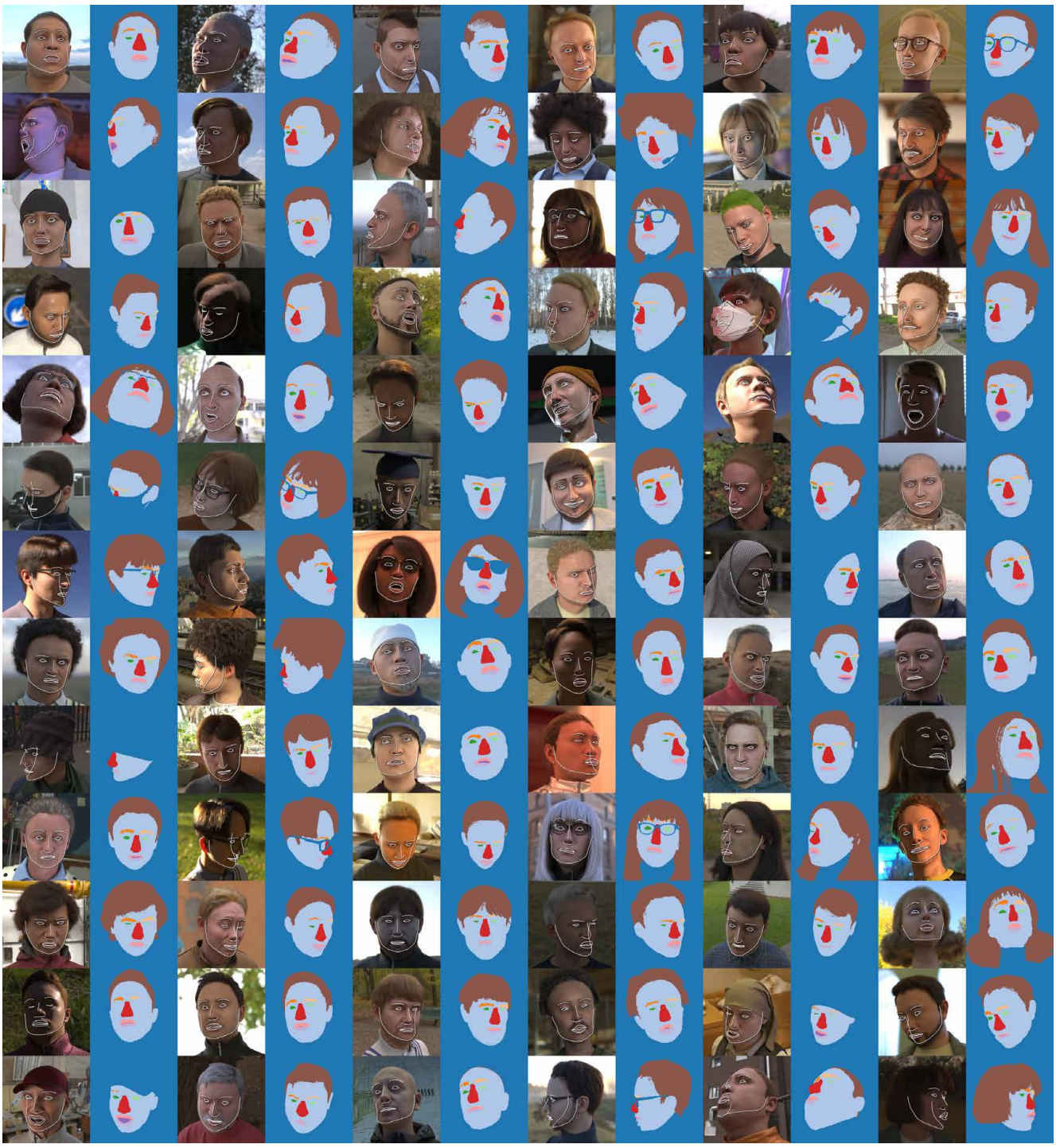

Figure 12. Here are some examples from the dataset we rendered, shown alongside the synthetic ground truth labels for landmark localization and face parsing. Note the variability in illumination, pose, identity, expression, camera location and focal length.

Figure 13. As we sample each of the assets independently, some renders can result in unlikely (but not impossible) combination of face shape, texture, hair style, make-up and accessories. However, we do not yet have any evidence that these unlikely faces actually harm machine learning training, and indeed may contribute to robustness.