Supplementary Material: Simulated Adversarial Testing of Face Recognition Models

Simulated Adversarial Testing Exploration

In order to explore samples generated by simulated adversarial testing and other simulated testing techniques, we are able to project the shape and texture components of our samples onto a plane of two components. We do so for the first two shape components (roughly controlling for height and width of the face). We show the results for adversarial testing, random optimization, Gaussian random testing and uniform random testing in Figure 1. First we observe the relative abundance of adversarial examples found using our method compared to other methods. Next, we can observe that adversarial testing not only finds adversarial examples, but that these examples are also very varied. We note that most unsuccessful runs of adversarial testing occur when the algorithm converges to local maxima that are located at the edges of the feasibility domain of $[-3, 3]$. All plots are drawn from samples tested on an ArcFace IR-SE-CBAM-ResNet50.

We now show two plots for adversarial testing on this network where we project the samples on the plane generated by the 1st and 2nd shape components, and the 3rd and 4th shape components. Here we discover an interesting phenomenon. In the 1st-2nd shape component plane, samples are varied and seem roughly uniformly distributed in the space. This means that although the 1st and 2nd shape component clearly have a role in finding adversarial samples, adversarial samples can be found with many different 1st and 2nd shape component values. The second plot shows that for the 3rd and 4th shape components, our adversarial sampling method clearly favors/disfavors some pockets in the space. For example we find that samples with average 3rd and 4th shape components tend not to be found by adversarial testing. We can test the hypothesis of whether these values are non adversarial by higher-dimensional grid search, although this would be time consuming. Another idea is to limit the feasibility domain to these average 3rd and 4th shape components, and run many instances of adversarial testing. If few or no adversarial samples are found then there is a chance that this is a space that is highly non-adversarial. We believe these types of projections can give a strong intuition over what features affect network performance. In Figure 3 and 4 we show the first four shape component variations for the FLAME model in frontal and profile poses. We can see that the 3rd and 4th shape component variations, while not overly noticeable in the frontal pose, introduce features that are only visible in the profile pose. For example when the 4th shape component is varied in the positive direction it tucks the subject's jaw in. This introduces a frontal/profile ambiguity, and the face verification network has a harder time correctly verifying pairs of these faces since it takes as input both the frontal and profile face images. Similarly, the 3rd shape component introduces a protuberance of the subject's head when varied in the positive direction, which is much more apparent in the profile image. This is congruent to the adversarial faces that we obtain in Figure 3 of the main paper that show frontal/profile ambiguous features.

Limitations and Future Work

Even though our adversarial testing algorithm using reinforcement learning is much more effective than random optimization and other sampling methods, it does not have a perfect rate of finding adversarial examples. It sometimes converges to local maxima that are hard to classify but nonetheless non-adversarial. We believe that one of the weaknesses is that it can get stuck in the boundary limits for the parameters that are being varied, and it has a hard time getting out of that space. This is especially a large problem in higher dimensional spaces. We will investigate this weakness in future work.

While we do exhibit for the first time the fact that two face recognition networks trained on the same data with different architectures and losses have vastly different loss landscapes when face shape and texture are varied, and thus are learning different things, we have not yet found suitable hypotheses that are verified by data that explain *why* this is the case. We believe this phenomenon requires more in-depth research and are working on verifying some of our hypotheses for our next work.

One of the major limitations of our work is the fact that we find adversarial examples that are simulated. We believe that the most challenging aspect of this research direction is verification of simulated adversarial samples using real data. This aspect is so challenging that it has been ne-

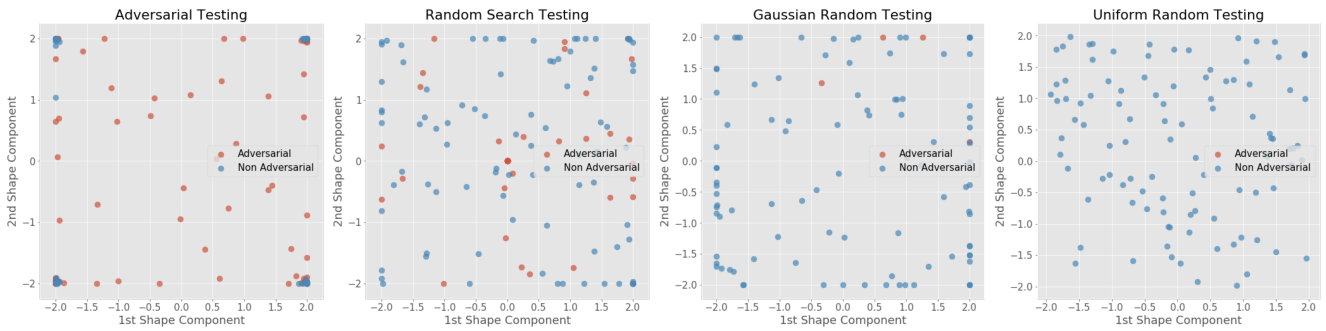

Figure 1. Comparison of adversarial/non-adversarial samples for different testing methods, projected onto the 1st and 2nd shape component plane.

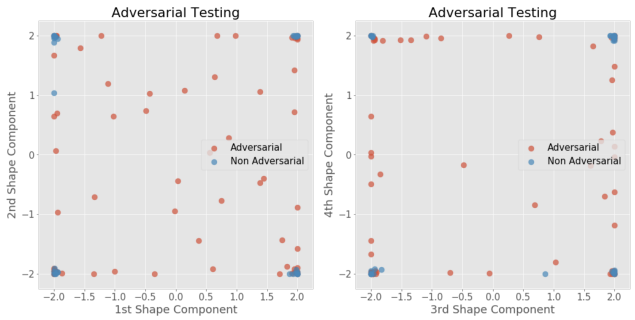

Figure 2. Comparison of adversarial/non-adversarial samples for simulated adversarial testing, projected onto the 1st/2nd and 3rd/4th shape component planes.

glected by prior research. Nevertheless, many works have been accepted to top conferences without treating this specific problem due to the potential they have in furthering our knowledge about robustness issues of neural networks with realistic variation of stimuli [1, 6, 9, 11, 12]. Even the best simulators that are currently available to the computer vision research community exhibit a substantial domain gap with real data [3, 7, 8, 10]. For this reason, it is difficult to verify the transfer capability of certain features. Of the different attributes that were available to us, head pose is one of the most reliable in terms of transfer due to several reasons: the ability to easily extract it from real images using a head pose estimation network, the 1-to-1 correspondence between head pose in simulated and real situations, and the low-dimensional nature of the attribute that can be more easily analyzed and plotted in a curve as shown in Figure 2. Due to all of these reasons we present the first link between simulated adversarial samples in the simulated and real world using this attribute. In the near-future, with more advanced simulators, we expect work to be able to confirm many more strong links between simulated and real samples. Just as [1] found that camera pose influences the predictions of a neural network in simulated data, we show

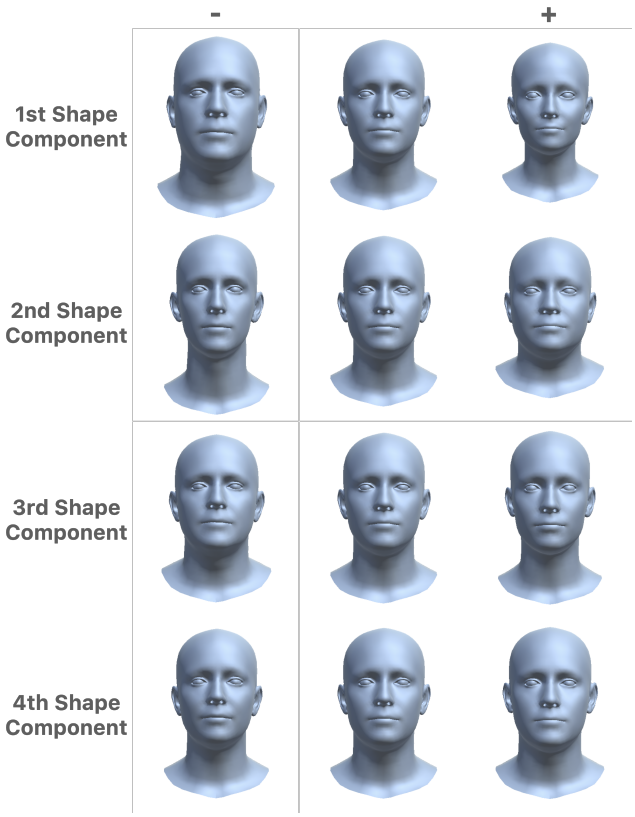

Figure 3. Shape variations along different shape components for the FLAME model in a frontal pose.

that pose, shape and texture jointly influence predictions of a face recognition network, but we go one step further and show that pose similarly impacts performance in the real and simulated world. Finally, in principle, it is almost impossible to find a real face that is arbitrarily close to any face we simulate. This is simply due to the fact that shape and texture are very high-dimensional, such that a point has very few close neighbors given a fixed-size real dataset. To find a very close sample we would have to collect a dataset

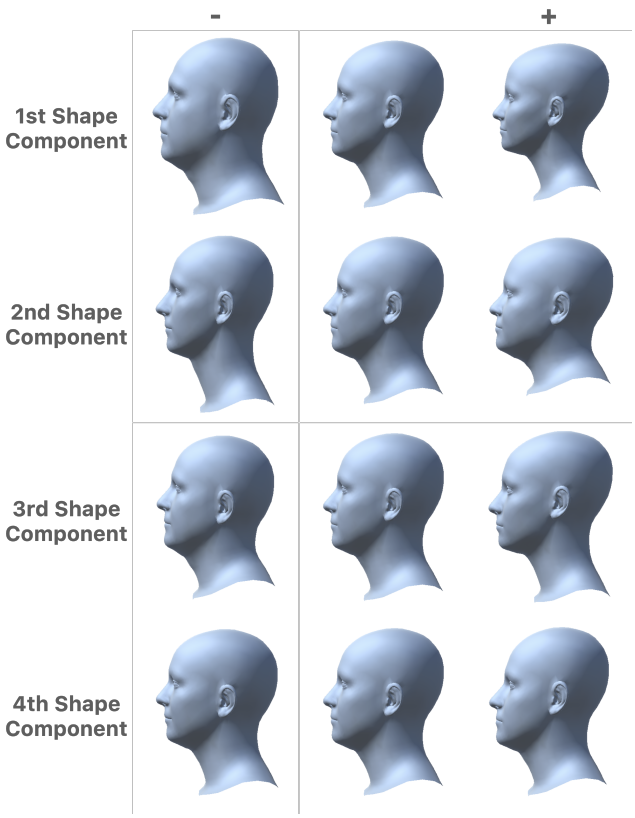

Figure 4. Shape variations along different shape components for the FLAME model in a profile pose.

that is extremely large.

Societal Impact

The plausible negative social consequences of this work are tightly linked with overall negative consequences of facial analysis systems. An approach that improves testing for face recognition systems such as the one we propose can be used to improve recognition rates on minorities, persecuted groups and oppressed individuals. This is a larger problem acting on any work that can potentially impact facial analysis, and we argue that our work has an asymmetric potential for applications that have a positive social impact. Given that researchers have proven that there exists gender and racial bias of beneficial face analysis systems [2, 4, 5], by better testing such systems these biases can be diagnosed and mitigated, meaning that minorities can more readily benefit from these technologies.

Another important point is that a major bottleneck for our work is a simulator that is expressive and realistic. Bias and lack of expressiveness of a simulator might mean that bias in the face recognition network is not correctly detected. We urge developers of future simulators to take into account the bias of their training population in order to increase the expressiveness of their simulator and decrease the bias. We also urge them to understand the power of such a tool for robustness and bias analysis and to distribute the model responsibly, similar to the FLAME head model [7] team.

References

[1] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019. 2

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. 3

[3] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2

[4] R. V. Garcia, L. Wandzik, L. Grabner, and J. Krueger. The harms of demographic bias in deep face recognition research. In *2019 International Conference on Biometrics (ICB)*, pages 1–6, 2019. 3

[5] Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. Face recognition vendor test part 3: Demographic effects. 2019. 3

[6] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2093–2102, 2018. 2

[7] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 3

[8] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[9] Nataniel Ruiz, Barry-John Theobald, Anurag Ranjan, Ahmed Hussein Abdelaziz, and Nicholas Apostoloff. Morphgan: One-shot face synthesis gan for detecting recognition bias. In *32nd British Machine Vision Conference 2021, BMVC 2021, Virtual Event, UK*, 2021. 2

[10] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics*, pages 621–635. Springer, 2018. 2

[11] Michelle Shu, Chenxi Liu, Weichao Qiu, and Alan Yuille. Identifying model weakness with adversarial examiner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11998–12006, 2020. 2

[12] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4302–4311, 2019. 2