UNIVERSITY OF NOTRE DAME
College of Engineering
COMPUTER VISION @ND

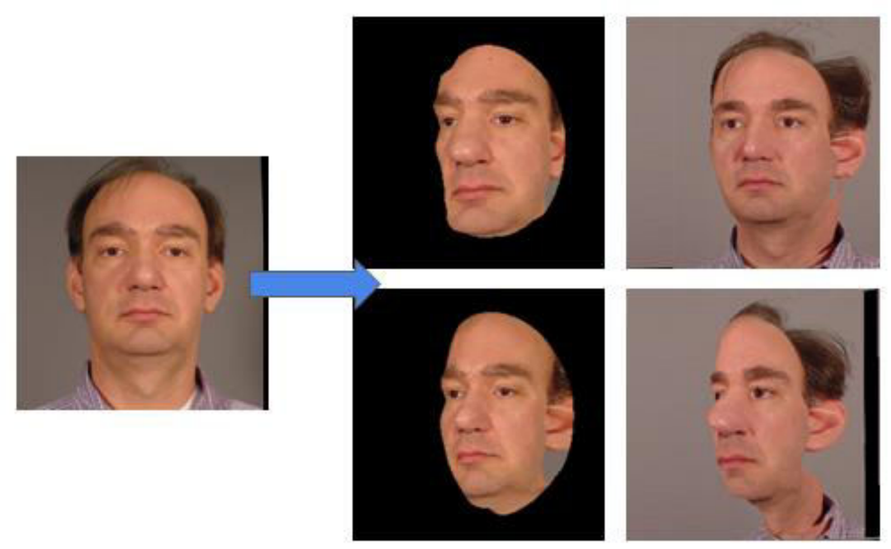
On Hallucinating Context & Background Pixels from a Face Mask using Multi-scale GANs

Sandipan Banerjee[1], Walter Scheirer[2], Kevin Bowyer[2], and Patrick Flynn[2]
[1]Affectiva, [2]University of Notre Dame, USA

WACV 2020

A. Motivation

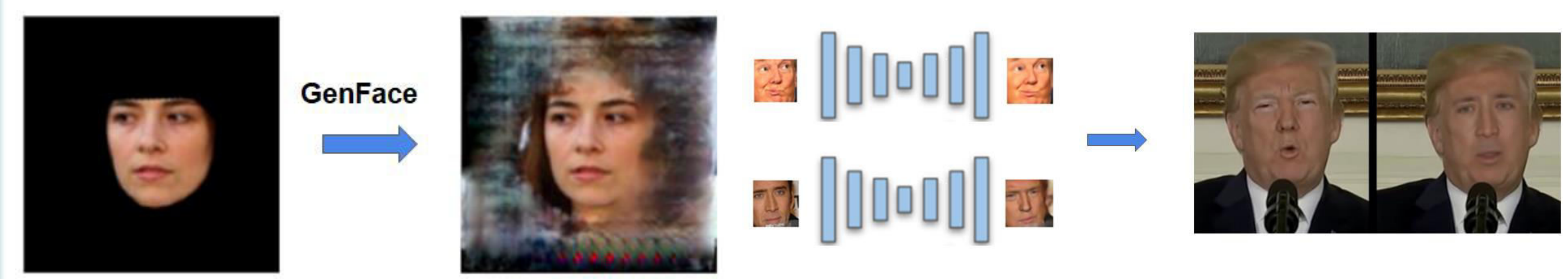
- Research has shown **presence of facial context** (forehead, hair, ears, neck, background) to **improve face verification** performance for both **human raters**[1] and **deep learning models**[2].
- Presence **of caps, hats, headgear, hood can occlude the facial context**, hampering verification accuracy[3], especially in surveillance videos.
- **Novel views** generated using 3D models do not have facial context[4] or do not look realistic[5].
- **Synthetic identities** generated using 3D models also have no facial context[6].



[1]Rice et al. Psychological Science 2013
[2]Phillips et al. FG 2017
[3]Singh et al. TBIOM 2019
[4,5]Masi et al. ECCV 2016, FG 2017
[6]Banerjee et al. WACV 2019

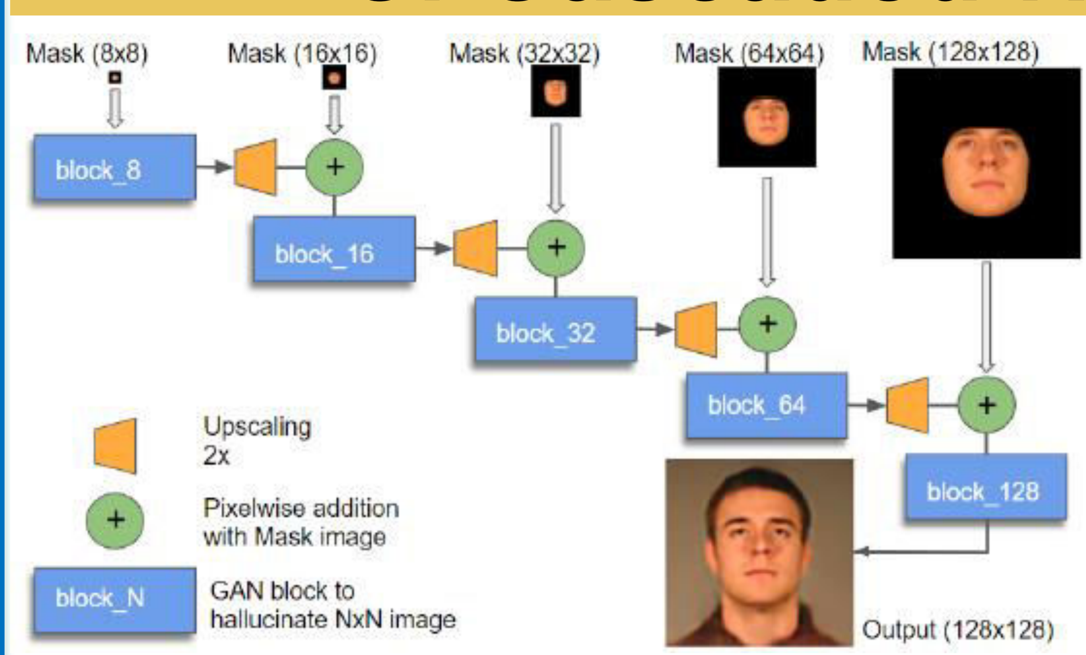
B. Existing Algorithms

- **Semantic Inpainting**[1] – These algorithms are primarily trained on hallucinate missing pixels in and around the facial mask. Hence they are not well suited for the task of synthesizing the whole facial context and background given a face mask, when used off-the-shelf pre-trained versions.
- **Face Swapping** - Strong similarity between source and target faces in terms of facial pose, ethnicity and gender[2]. Even then identity is not well preserved[3].
- **DeepFake**[4,5] – Subject specific autoencoder models, not applicable for real world applications.

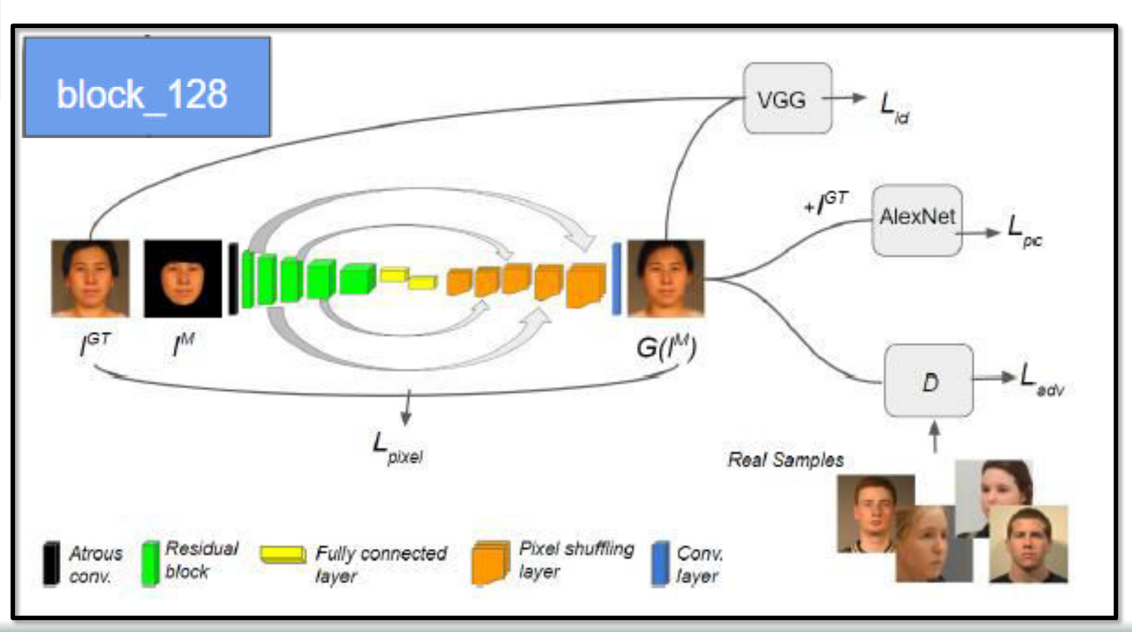


[1]Li et al. CVPR 2017
[2]Kemelmacher-Shlizerman et al. SIGGRAPH 2016
[3]Nirkin et al. ICCV 2019
[4]Korshunova et al. ICCV 2017
[5]https://github.com/deepfakes/faceswap

C. Cascaded Network of GANs

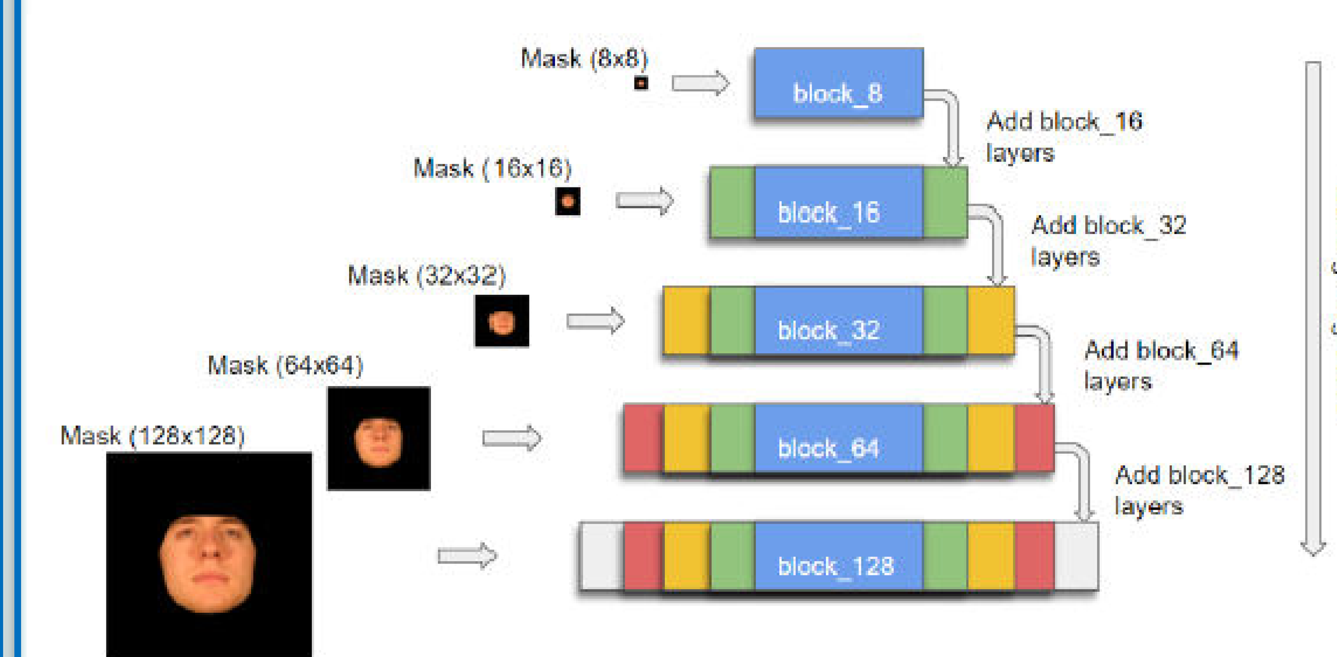


$$L = L_{pixel} + \lambda_1 L_{pc} + \lambda_2 L_{adv} + \lambda_3 L_{id} + \lambda_4 L_{tv}$$



Ulyanov et al. ICML 2016, Shi et al. CVPR 2016
Parkhi et al, BMVC 2015, Zhang et al. CVPR 2018

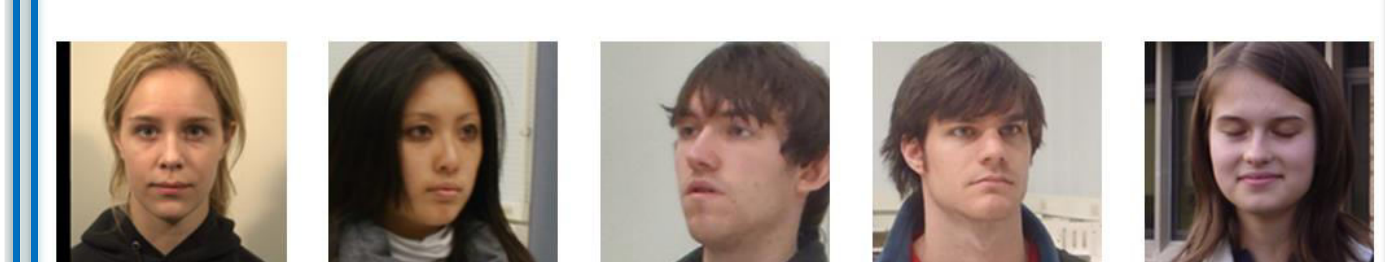
D. Progressive Growing



Karras et al. ICLR 2018,

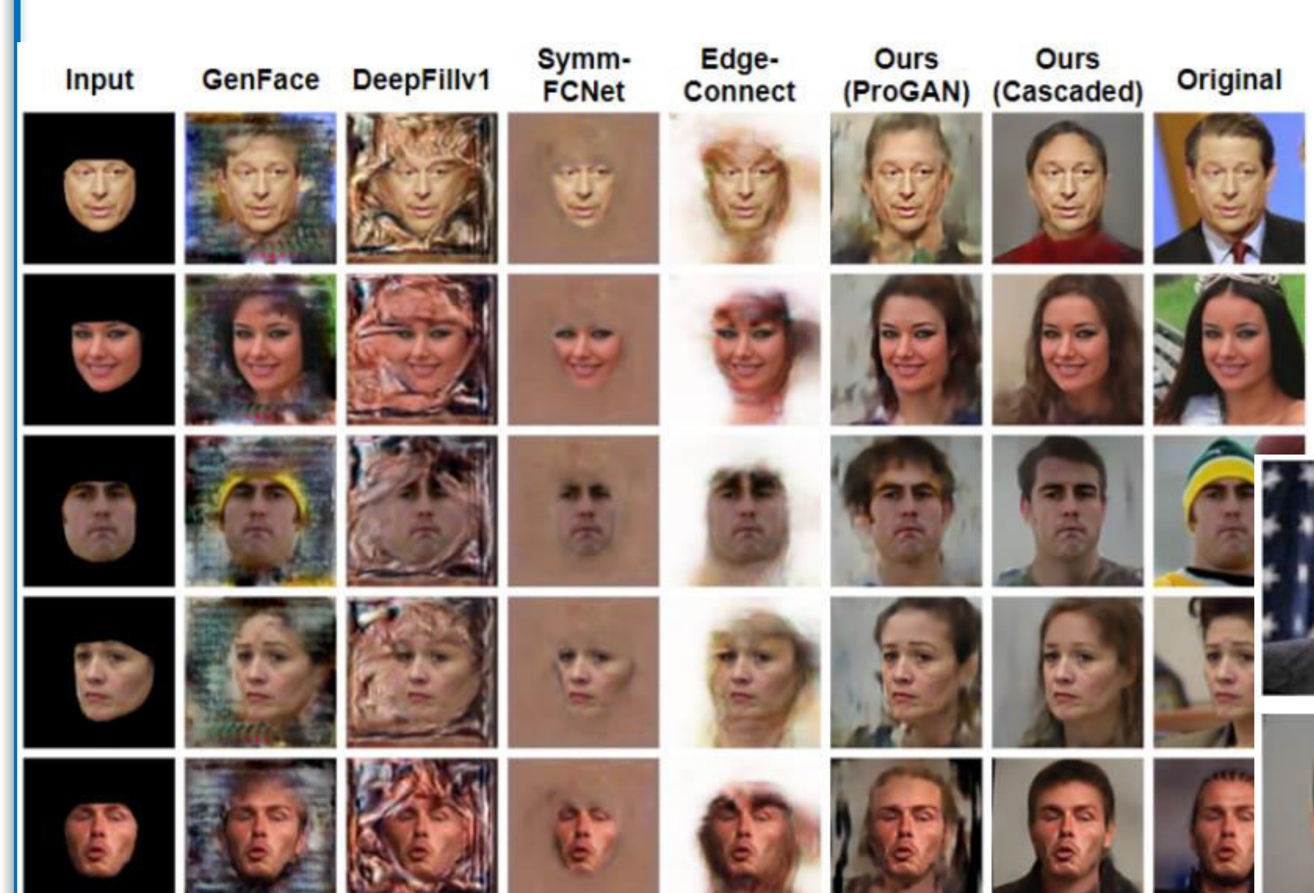
E. Training Data

- For training our model, we randomly sample only **12,000 face** images (7,495 male and 4,595 female)[1].
- **Image mirroring** is applied to augment the available poses in the training set.
- Masked faces obtained using a recent pre-trained face alignment model[2] resized to 128x128x3.



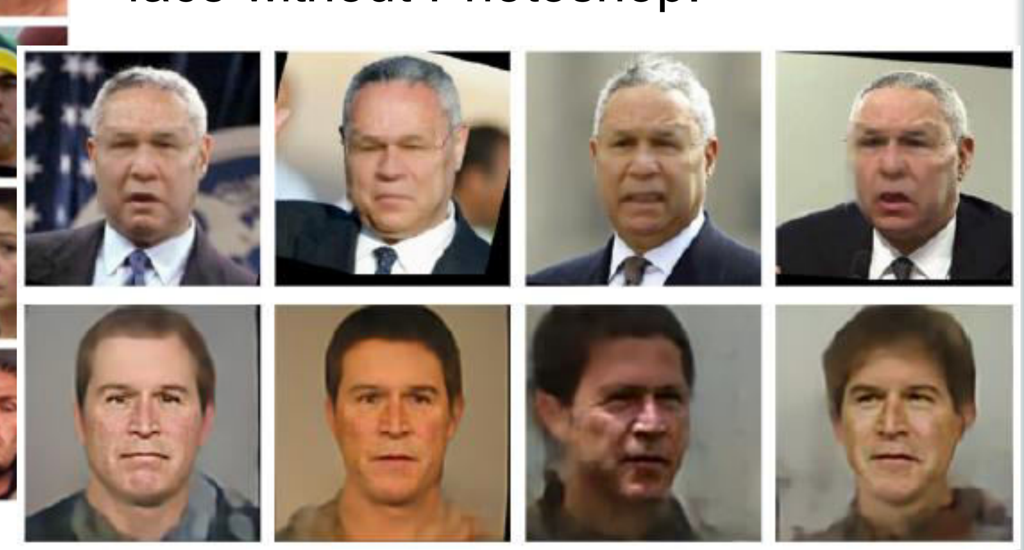
[1]Phillips et al. Image and Vision Computing 2016
[2]Bulat et al. ICCV 2017

F. Qualitative Results



GenFace - Li et al. CVPR 2017, DeepFillv1 - Yu et al. CVPR 2018
SymmFCNet – Li et al. arXiv:1812.07741,
EdgeConnect – Nazeri et al. ICCV Workshops 2019

- DeepFake can be **trained on only one source subject** at a time
- **Blending seam visible** across the face without Photoshop.



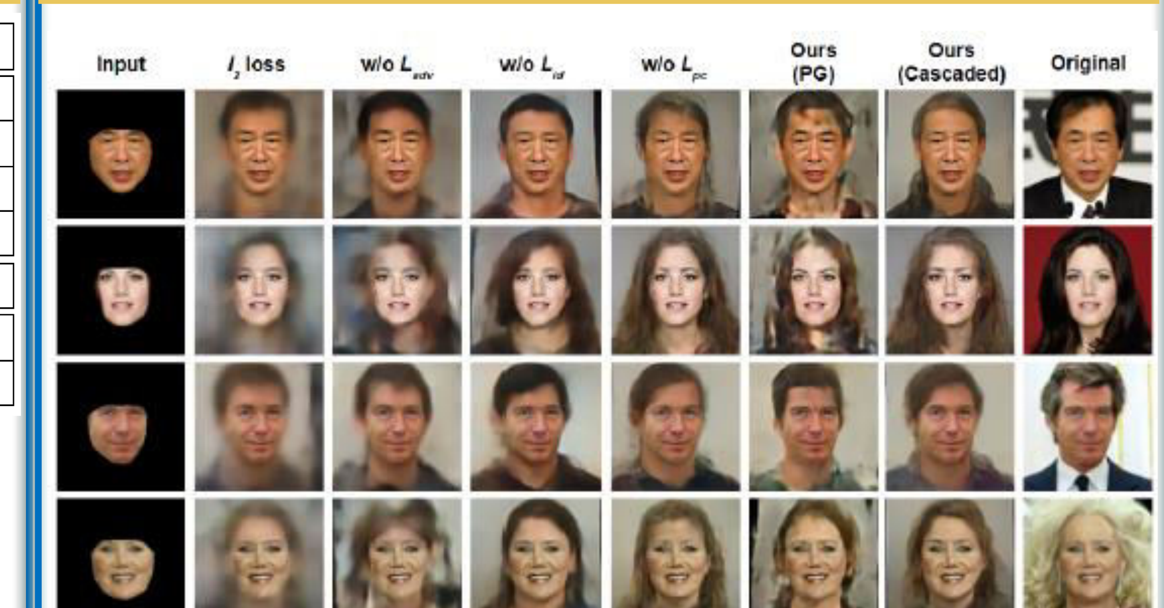
G. Quantitative Results

Model	Mean Match Score	Mean SSIM [70]	FID [26]	Mean Perceptual Error [61]
GenFace [44]	0.543	0.491	177.06	3.536
DeepFillv1 [73]	0.481	0.321	241.696	3.204
SymmFCNet [43]	0.457	0.333	207.117	2.434
EdgeConnect [55]	0.454	0.178	141.695	3.106
DeepFake	0.459	0.448	**43.03**	1.857
Ours (ProGAN)	0.668	0.466	103.71	2.255
Ours (Cascaded)	**0.722**	**0.753**	46.12	**1.256**

Training Data	CW [74] Images (Identities)	Hallucinated Images (Identities)	LFW [27] Performance (TPR@FPR = 0.01)
Dataset 1	494,414 (10,575)	0	0.963
Dataset 2	494,414 (10,575)	494,414 (10,575)	0.971

FID – Heusel et al. NeurIPS 2017
Mean Perceptual Error - Prashnani et al. CVPR 2018
CW - Yi et al. arXiv:1411.7923
ResNet50 - He et al. CVPR 2016
LFW - Huang et al. Tech Report, 2007

H. Ablation Studies



Model	Mean Match Score	Mean SSIM [16]	FID [6]	Mean Perceptual Error [12]
l_2 loss	0.520	0.413	166.76	2.489
w/o L_{adv}	0.522	0.411	132.71	2.320
w/o L_{id}	0.609	0.519	91.65	1.956
w/o L_{pc}	0.624	0.528	101.44	2.046
Ours (ProGAN)	0.668	0.466	103.71	2.255
Ours (Cascaded)	**0.722**	**0.753**	**46.12**	**1.256**

* Hardware support was generously provided by NVIDIA.