Leveraging Diffusion For Strong and High Quality Face Morphing Attacks

Zander W. Blasingame and Chen Liu

Department of Electrical and Computer Engineering
Clarkson University
{blasinzw, cliu}@clarkson.edu

Read Our Paper!

Motivation

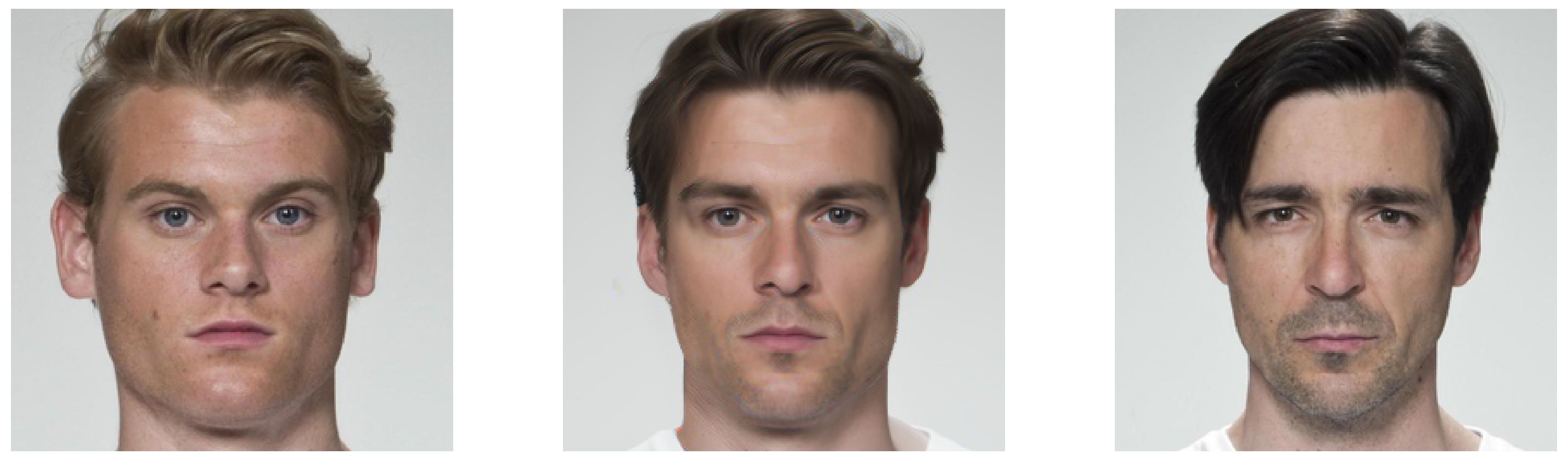

(a) Identity a (b) Morphed image (c) Identity b

Figure 1. Example of the proposed Diffusion-based morphing attack. Samples are from FRLL dataset.

- Face Recognition (FR) systems are vulnerable to face morphing attacks [1].
- Two classes of morphing attacks: **landmark-based** attacks and **deep learning-based** attacks.
- Nearly all state-of-the-art deep learning-based attacks are based on the GAN framework.
- Diffusion-based methods have been shown to outperform GANs [2].

Methodology

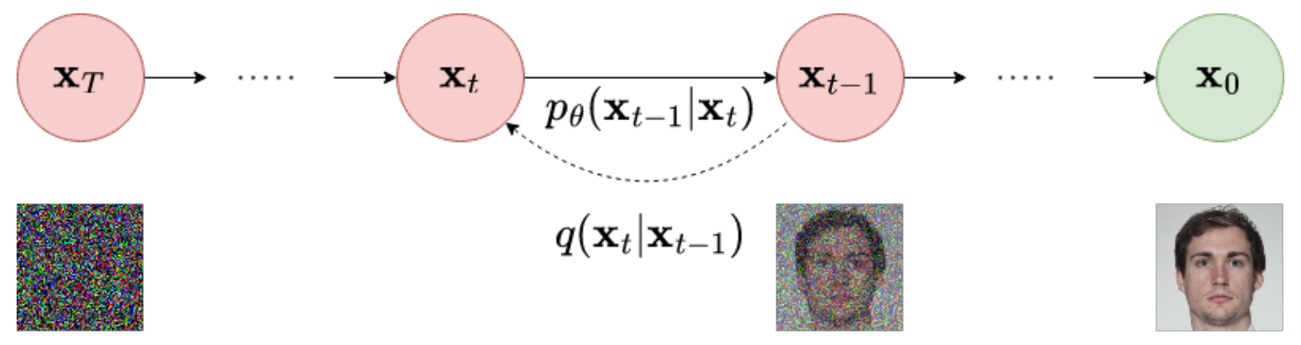

$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$\mathbf{x}_0$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$

Figure 2. The forward and reverse Diffusion processes.

- Diffusion method gradually destroys an image by adding noise, $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$.
- Learn reverse trajectory $p_\theta(\mathbf{x}_{0:T})$ by optimizing the evidence lower bound (ELBO).
- Using the Denoising Diffusion Implicit Model (DDIM) scheduler allows for deterministic generation

$$\mathbf{x}_{t-1} = \frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}}(\mathbf{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta^{(t)}(\mathbf{x}_t)) + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_\theta^{(t)}(\mathbf{x}_t) \tag{1}$$

where $\boldsymbol{\epsilon}_\theta^{(t)}$ is a learned noise predictor and $\alpha_t = 1 - \beta_t$ for variance schedule $\{\beta_t\}_{t=1}^T$.
- Diffusion autoencoders embed both stochastic and semantic details in twin latent spaces [3].
- Condition forward and reverse trajectories on latent embedding $\mathbf{z} = E(\mathbf{x}_0)$.
- Training loss is done via a simplified loss function.

$$L = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_0 \sim q(\mathbf{x}_0) \\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0},\mathbf{I})}} \|\boldsymbol{\epsilon}_\theta^{(t)}(\mathbf{x}_t, \mathbf{z}) - \boldsymbol{\epsilon}_t\|_2^2 \tag{2}$$

- Morph both stochastic $\ell_{\mathcal{X}}(\mathbf{x}_T^{(a)}, \mathbf{x}_T^{(b)}; 0.5)$ and semantic $\ell_{\mathcal{Z}}(\mathbf{z}_a, \mathbf{z}_b; 0.5)$ latent codes.
- Stochastic interpolation is spherical, semantic is linear.
- Preform rudimentary “pre-morph” in image space $\xi(\mathbf{x}_0^{(a)}, \mathbf{x}_0^{(b)})$ before diffusing.
- Morphed semantic latent guides generative process.

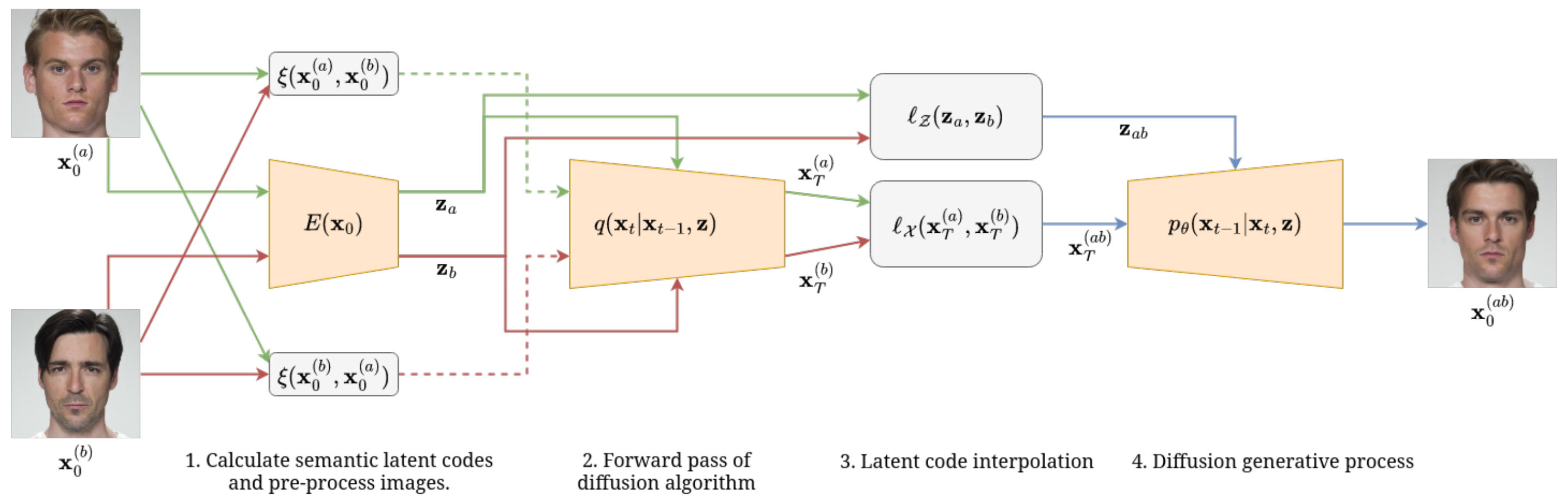

$\mathbf{x}_0^{(a)}$
$\mathbf{x}_0^{(b)}$
$\xi(\mathbf{x}_0^{(a)}, \mathbf{x}_0^{(b)})$
$\xi(\mathbf{x}_0^{(b)}, \mathbf{x}_0^{(a)})$
$E(\mathbf{x}_0)$
$\mathbf{z}_a$
$\mathbf{z}_b$
$q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z})$
$\mathbf{x}_T^{(a)}$
$\mathbf{x}_T^{(b)}$
$\ell_{\mathcal{Z}}(\mathbf{z}_a, \mathbf{z}_b)$
$\ell_{\mathcal{X}}(\mathbf{x}_T^{(a)}, \mathbf{x}_T^{(b)})$
$\mathbf{z}_{ab}$
$\mathbf{x}_T^{(ab)}$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z})$
$\mathbf{x}_0^{(ab)}$
1. Calculate semantic latent codes and pre-process images.
2. Forward pass of diffusion algorithm
3. Latent code interpolation
4. Diffusion generative process

Figure 3. Proposed architecture for Diffusion-based morphs, where the green traces indicate variables associated with identity a, likewise red traces denote identity b, and blue traces for the morphed identity ab.

Experimental Setup

- FERET [4], FRLL [5], and FRGC v2.0 [6] datasets were used to evaluate the proposed attack.
- Evaluated performance against two publicly available state-of-the-art face recognition systems: FaceNet and VGGFace2.
- Compared against four other morphing attacks: OpenCV, FaceMorpher, StyleGAN2, and MIPGAN-II.
- OpenCV and FaceMorpher are landmark-based attacks.
- StyleGAN2 and MIPGAN-II are based on the StyleGAN2 architecture.
- The OpenCV, FaceMorpher, and StyleGAN2 morphed images were generated by [7].
- The MIPGAN-II morphs were created by [8].

Evaluation of Visual Fidelity

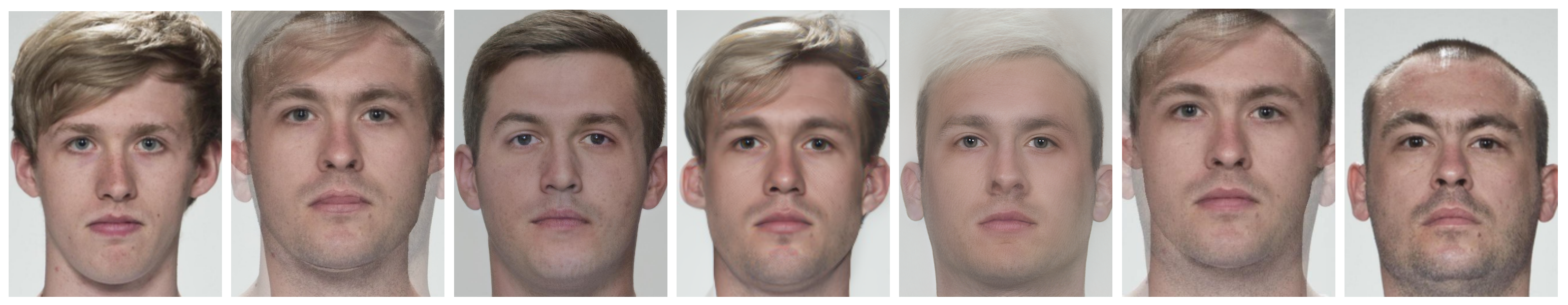

(a) Identity a (b) OpenCV (c) StyleGAN2 (d) Diffusion (e) MIPGAN-II (f) FaceMorph (g) Identity b

Figure 4. Different generated morphs from two identities from the FRLL dataset.

- The visual fidelity is measured using the Fréchet Inception Distance (FID).
- The FID is defined as the Fréchet (2-Wasserstein) distance between the activations of the deepest layer of the Inception v3 network.
- The 2-Wasserstein metric between two probability measures μ, ν with finite moments on $\mathbb{R}^n$ is defined as

$$W_2(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2^2 \, \mathrm{d}\pi(x, y) \right)^{\frac{1}{2}} \tag{3}$$

where $\Pi(\mu, \nu)$ is the set of all distributions with marginals μ and ν.
- The FID is measured between the morphed images and genuine images for each dataset.

Table 1. FID across different morphing attacks. Lower is better.

Morphing Attack	FRLL	FRGC	FERET
StyleGAN2	45.19	86.41	**41.91**
FaceMorpher	91.97	88.14	79.58
OpenCV	85.71	100.02	91.94
MIPGAN-II	66.41	115.96	70.88
Diffusion	**42.63**	**64.16**	50.45

Vulnerability of FR Systems

- The Mated Matched Presentation Match Rate (MMPMR), specifically the ProdAvg-MMPMR variant, is used to evaluate the vulnerability of an FR system to a morphing attack.

Table 2. MMPMR at FMR = 0.1% across different morphing attacks. Higher is better.

	FRLL		FRGC		FERET		
Morphing Attack	FaceNet	VGGFace2	FaceNet	VGGFace2	FaceNet	VGGFace2	Geometric Mean
StyleGAN2	4.69	6.05	0.18	0.85	0.54	0.76	1.10
FaceMorpher	11.26	36.4	0.51	9.15	2.3	10.78	6.02
OpenCV	17.34	40.93	0.14	12.16	1.69	11.12	5.32
MIPGAN-II	30.96	26.74	3.12	7.94	6	5.39	9.34
Diffusion	28.14	35.37	2.68	8.47	6.47	13.03	**11.13**

Detectability of Morphing Attacks

Table 3. Ablation study on the impact morphing attack on validation accuracy.

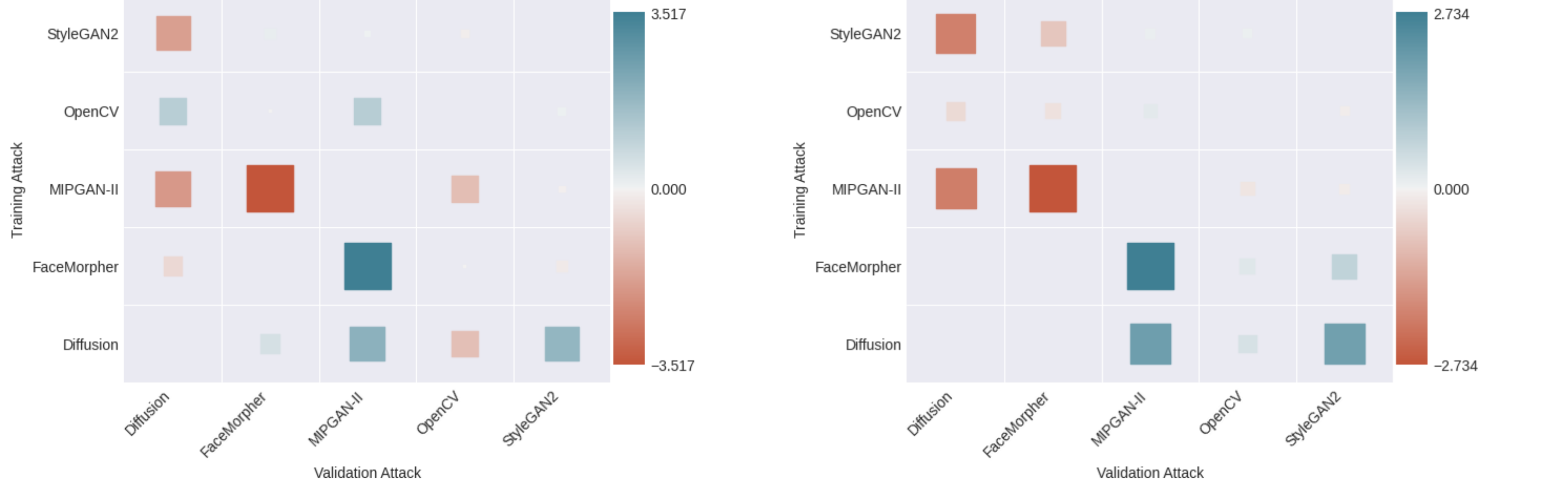
	Training Attack					Validation Attack				
Dataset	Diffusion	FaceMorpher	MIPGAN-II	OpenCV	StyleGAN2	Diffusion	FaceMorpher	MIPGAN-II	OpenCV	StyleGAN2
FERET	✗	✓	✓	✓	✓	72.73	99.23	100	99.95	99.33
FERET	✓	✗	✓	✓	✓	99.9	76.39	100	99.85	99.64
FERET	✓	✓	✗	✓	✓	99.69	99.38	100	99.95	99.54
FERET	✓	✓	✓	✗	✓	99.74	99.48	100	99.74	99.43
FERET	✓	✓	✓	✓	✗	99.74	98.56	99.9	99.74	87.89
FRGC	✗	✓	✓	✓	✓	75.89	99.98	99.97	99.9	99.93
FRGC	✓	✗	✓	✓	✓	99.95	99.48	100	99.9	99.95
FRGC	✓	✓	✗	✓	✓	99.83	99.85	99.82	99.8	99.85
FRGC	✓	✓	✓	✗	✓	99.93	100	100	99.23	99.93
FRGC	✓	✓	✓	✓	✗	99.93	99.93	99.94	99.88	97.83
FRLL	✗	✓	✓	✓	✓	13.96	99.58	99.32	99.65	99.65
FRLL	✓	✗	✓	✓	✓	99.23	99.09	98.91	99.37	99.44
FRLL	✓	✓	✗	✓	✓	99.09	98.95	98.24	99.02	99.09
FRLL	✓	✓	✓	✗	✓	99.51	99.44	99.19	99.16	99.58
FRLL	✓	✓	✓	✓	✗	99.93	99.86	99.86	99.93	95.02

- We propose a metric to measure the relative strength between morphing attacks.
- The transferability of morphing attack α to β is defined as

$$T(\alpha, \beta) = P(f^\alpha(X^\beta) = 1 \mid f^\alpha(X^\alpha) = 1) \tag{4}$$

where X^α, X^β are morphs created by α, β and f^α is a detector trained on α.
- The relative strength metric (RSM) from α to β is:

$$\Delta(\alpha \| \beta) = \log \left(\frac{T(\alpha, \beta)}{T(\beta, \alpha)} \right) \tag{5}$$



(a) RSM on FRGC (b) RSM on FERET

Figure 5. Blue indicates strong strength and red indicates weak strength.

Conclusion

- Novel state-the-of-the-art morphing attack with high visual fidelity.
- Diffusion morphs are able to fool FR systems while retaining high visual fidelity.
- Novel metric to compare the relative strength of morphing attacks.
- Diffusion morphs are very difficult to detect if the detector is not trained against them.

References

[1] Z. Blasingame and C. Liu, “Leveraging adversarial learning for the detection of morphing attacks,” *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–8, 2021.

[2] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 8780–8794, Curran Associates, Inc., 2021.

[3] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, “Diffusion autoencoders: Toward a meaningful and decodable representation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10619–10629, June 2022.

[4] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, “The feret database and evaluation procedure for face-recognition algorithms,” *Image Vis. Comput.*, vol. 16, pp. 295–306, 1998.

[5] L. DeBruine and B. Jones, “Face research lab london set.”

[6] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of the face recognition grand challenge,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 947–954 vol. 1, 2005.

[7] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, “Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks,” *ArXiv*, vol. abs/2012.05344, 2020.

[8] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, “Mipgan—generating strong and high quality morphing attacks using identity prior driven gan,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 365–383, 2021.

https://camel.clarkson.edu

EAB & CITeR Biometrics Workshop 2023, Martigny, Switzerland

Supported by CITeR and NSF under Grant # 1650503