ICLR

Vec2Face: Scaling Face Dataset Generation with Loosely Constrained Vectors

Australian
National
University

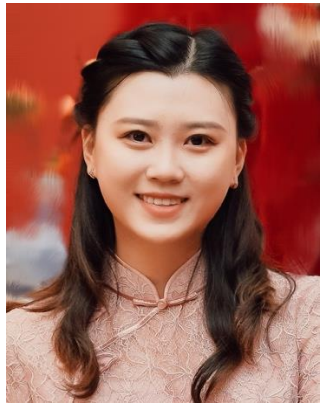
INDIANA UNIVERSITY
SOUTH BEND

Haiyu Wu

Uni. Notre Dame

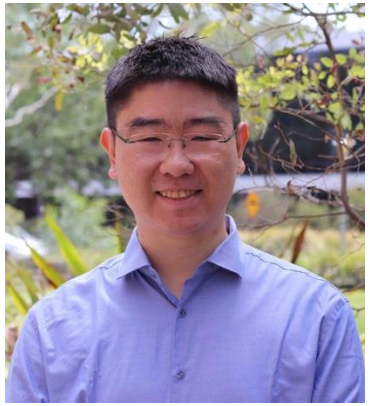
Jaskirat Singh

ANU



Sicong Tian

Indiana University



Liang Zheng

ANU

Kevin W. Bowyer

Uni. Notre Dame

Outline

- Background
- Proposed algorithm: Vec2Face
- Dataset assembling
- Performance and analysis

Background

Training set (real)

Model training

Test set (real)

General Protection Regulation (GDPR)
California Consumer Privacy Act (CCPA)
Act on the Protection of Personal Information (AAPI)...

Challenges

Training set (syn.)

Model training

Test set (real)

Realism – not much of an issue
We have VAE, GAN, Diffusion…

Large number of unique IDs / classes
Large intra-class variation

https://github.com/NVlabs/stylegan

ID separability in synthetic faces

SynFace (Qiu et al., 2021)

ID1:

ID2:

SFace (Boutros et al., 2022)

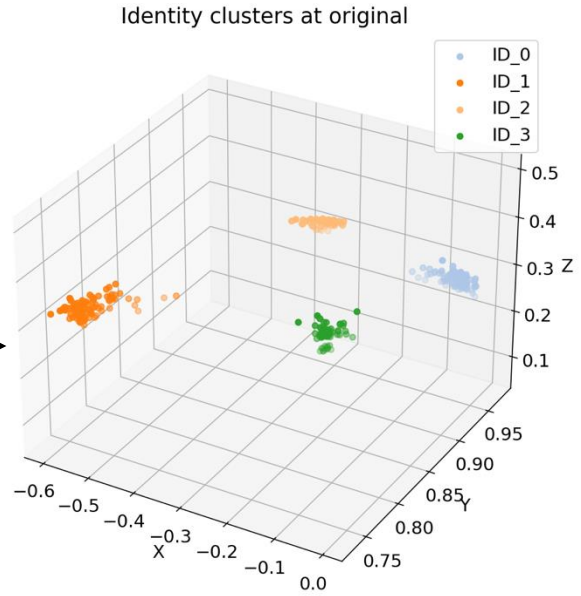
IDiff-Face (Boutros et al., 2023)

Intuition

Reverse

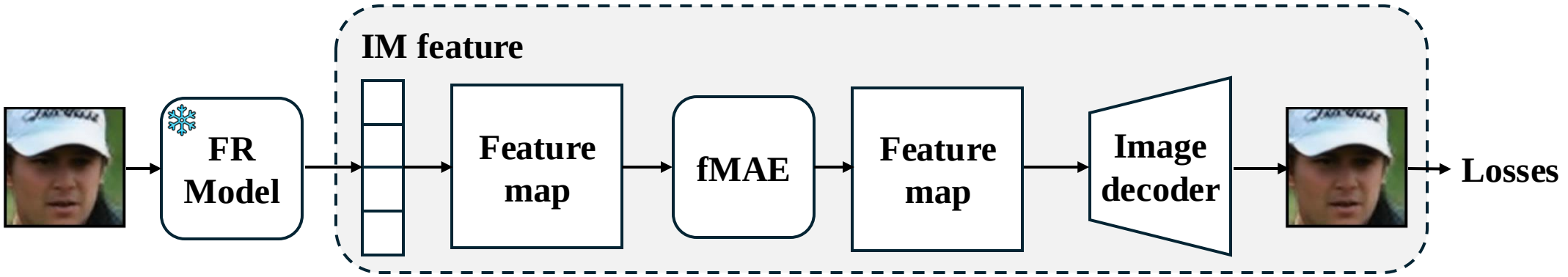
Face Recognition

$V^{512\text{-dim}}$

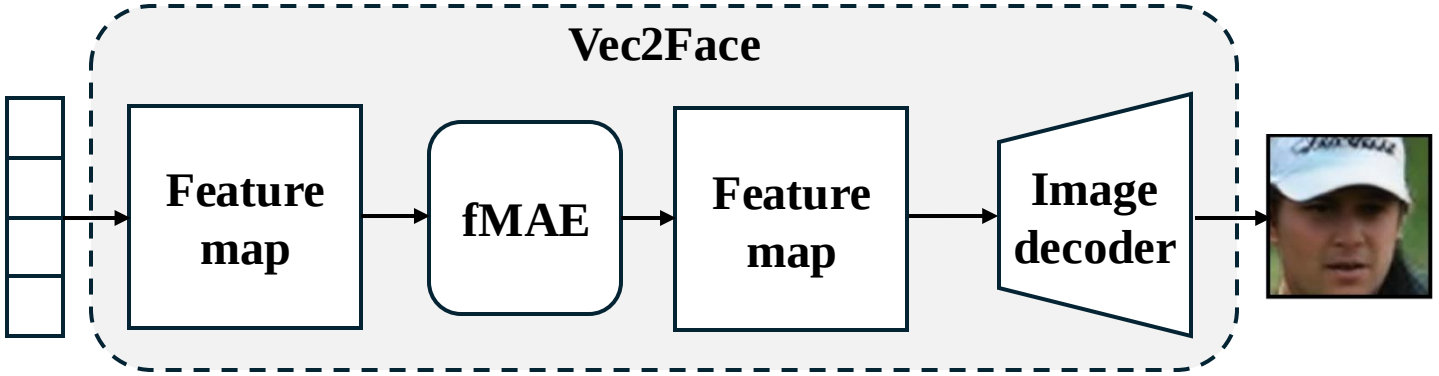


But image to image is the popular training paradigm...

Training



Inference



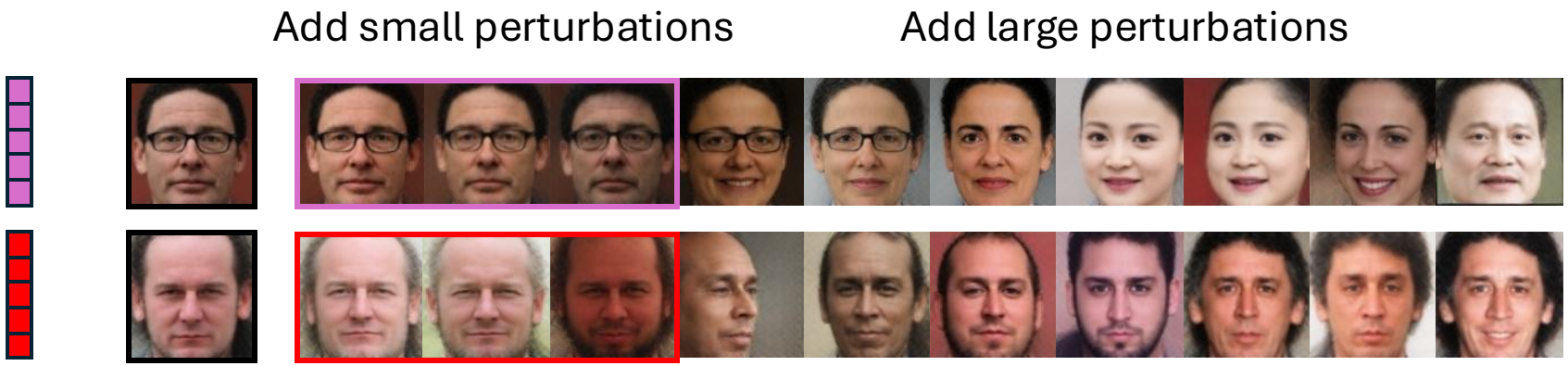
Wu et al., Vec2Face: Scaling face dataset generation with loosely constrained vectors. ICLR 2025

Properties

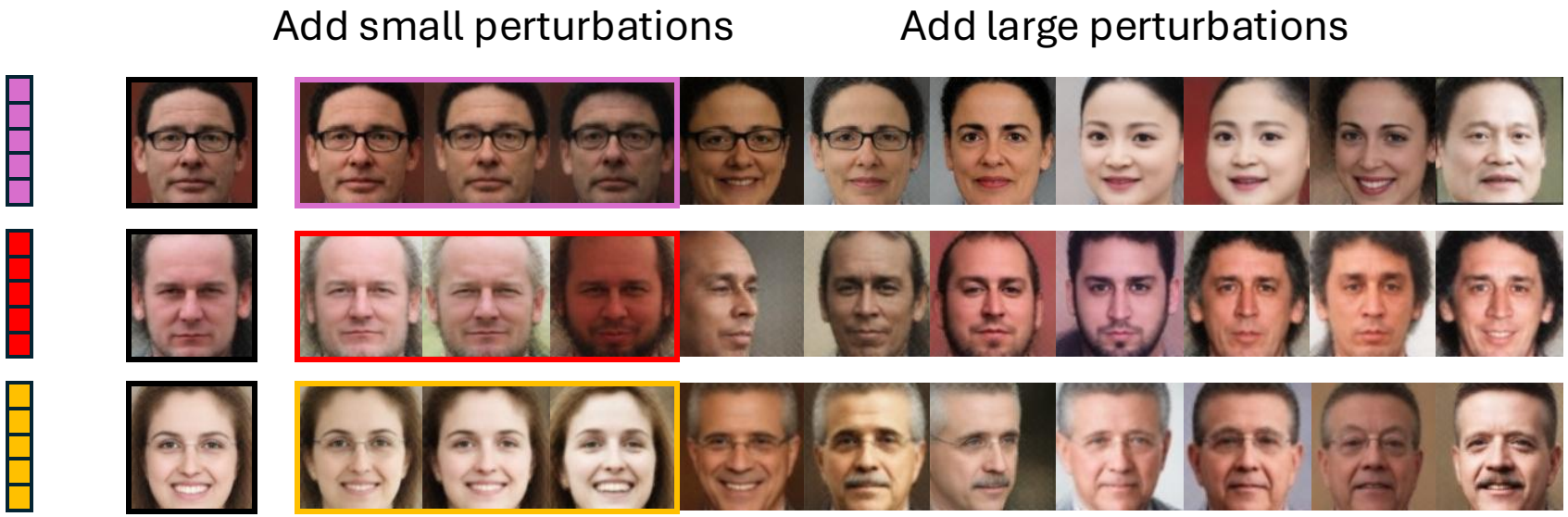
Add small perturbations

Add large perturbations

Properties



Properties

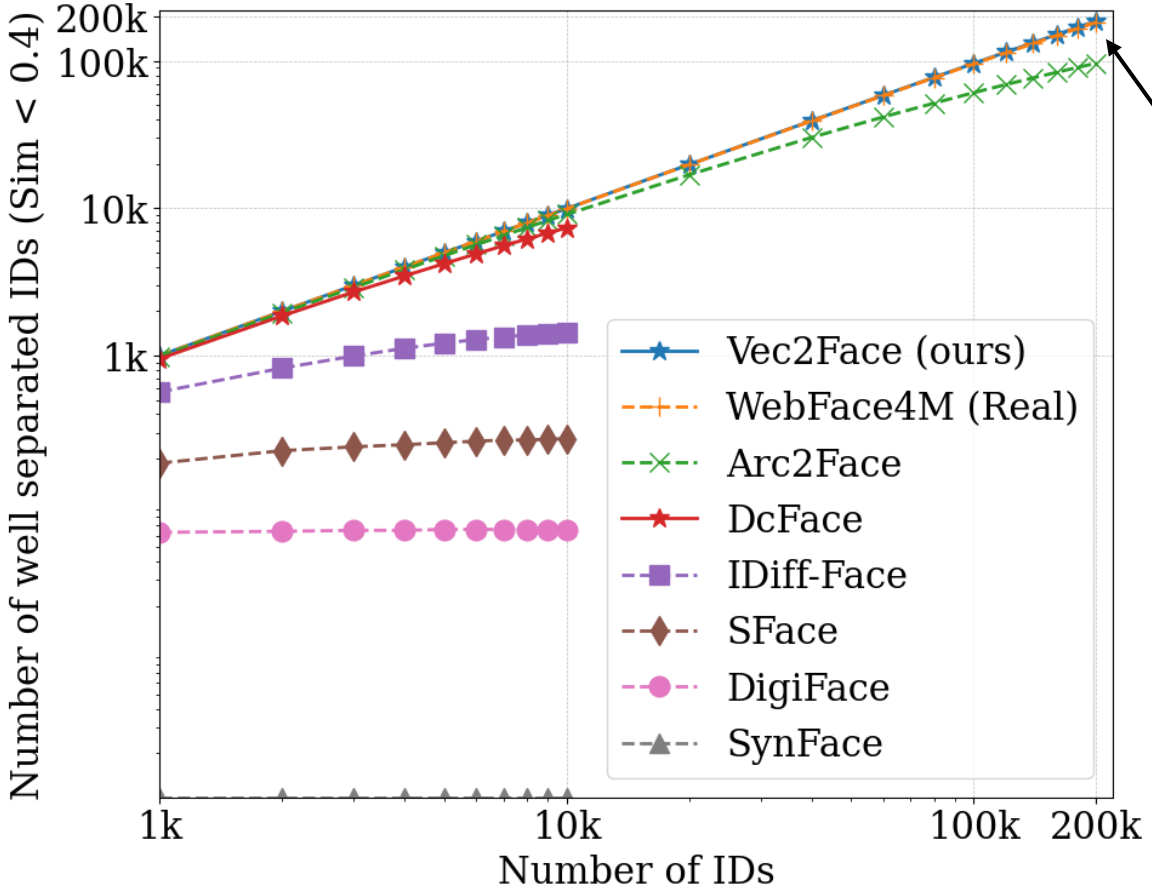


Well separated identities:

A vector creates a unique identity if its similarity to any other feature is less than a threshold.

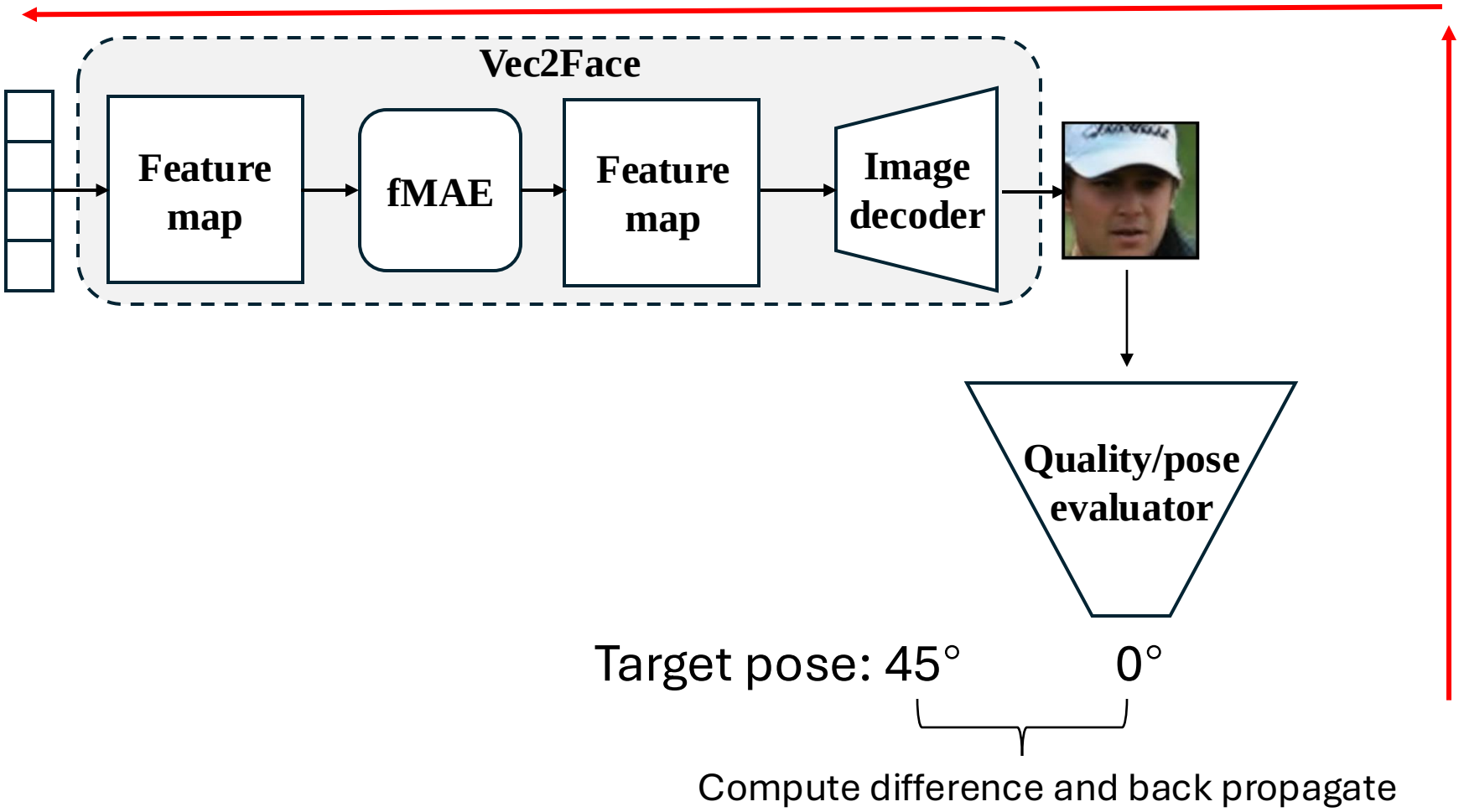
Wu et al., Vec2Face: Scaling face dataset generation with loosely constrained vectors. ICLR 2025

Identity separability



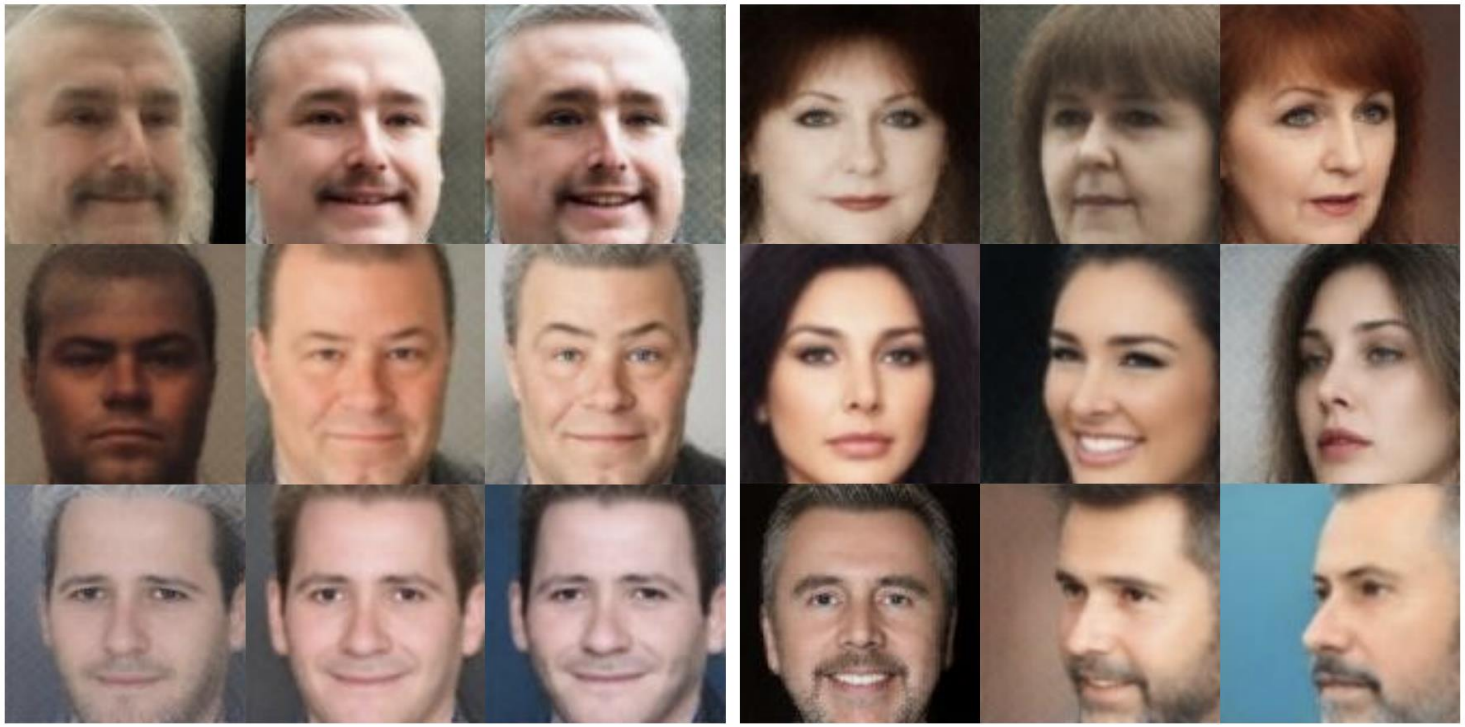
Our model can generate a large amount of well-separated identities

Attribute control with Vec2Face



Wu et al., Vec2Face: Scaling face dataset generation with loosely constrained vectors. ICLR 2025

Attribute control with Vec2Face



Quality control

Pose control

Wu et al., Vec2Face: Scaling face dataset generation with loosely constrained vectors. ICLR 2025

Training sets	# images	LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg.
IDiff-Face (Boutros et al., 2023a)†	0.5M	98.00	85.47	80.45	86.43	90.65	88.20
DCFace (Kim et al., 2023)†	0.5M	98.55	85.33	82.62	89.70	91.60	89.56
Arc2Face (Papantoniou et al., 2024)†	0.5M	98.81	**91.87**	85.16	90.18	92.63	91.73
DigiFace (Bae et al., 2023)⋆	1M	95.40	87.40	78.87	76.97	78.62	83.45
SynFace (Qiu et al., 2021)◇	0.5M	91.93	75.03	70.43	61.63	74.73	74.75
SFace (Boutros et al., 2022a)◇	0.6M	91.87	73.86	73.20	71.68	77.93	77.71
IDnet (Kolf et al., 2023)◇	0.5M	92.58	75.40	74.25	63.88	79.90	79.13
ExFaceGAN (Boutros et al., 2023b)◇	0.5M	93.50	73.84	71.60	78.92	82.98	80.17
SFace2 (Boutros et al., 2024)◇	0.6M	95.60	77.11	74.60	77.37	83.40	81.62
Langevin-Disco (Geissbühler et al., 2024)◇	0.6M	96.60	73.89	74.77	80.70	87.77	82.75
HSFace10K (Ours)◇	0.5M	**98.87**	88.97	**85.47**	**93.12**	**93.57**	**92.00**
CASIA-WebFace (Real)	0.49M	99.38	96.91	89.78	94.50	93.35	94.79

Accuracy on other test sets

Datasets	Hadrian	Eclipse	SLLFW	DoppelVer
HSFace10K	69.47	64.55	92.87	86.91
HSFace20K	75.22	67.55	94.37	88.90
HSFace100K	**80.00**	**70.35**	95.58	90.39
HSFace200K	**79.85**	**71.12**	95.70	89.86
HSFace300K	**81.55**	**71.35**	95.95	90.49
CASIA-WebFace	77.82	68.52	**96.95**	**95.11**

Table 6: Comparing a real dataset with HSFaces on other tasks. Hadrian, Eclipse, SLLFW, and DoppelVer emphasize facial hair variation, face exposure difference, similar-looking, and doppelganger, respectively.

Datasets	# of images	IJBB	IJBC
DCFace	0.5M	66.47	69.92
HSFace10K	0.5M	**83.82**	**86.96**
CASIA-WebFace	0.5M	78.71	83.44
HSFace100K	5M	86.16	89.73
WebFace4M	4M	**95.07**	**96.63**

Table 1: TPR@FPR=1e-4 on IJBB and IJBC datasets. The models are trained with SE-IResNet50.

Identity leakage

5M randomly sampled IDs	0.5	0.7
WebFace4M	0	0
CASIA-WebFace	0	0

Table 1: Identity leakage experiment between 5M randomly sampled identities and real identities. According to WebFace4M Zhu et al. (2023), identity pairs with similarity larger than 0.7 can be regarded as the same identity. We report the percentage of identity pairs falling in this range.

Takeaways

- Our model can generate large number of well-separated IDs w/o ID leakage.
- Our algorithm can accurately control the facial attributes.
- The generated training sets result in better performance than a real training dataset on various tasks.

ICLR

Vec2Face: Scaling Face Dataset Generation with Loosely Constrained Vectors

Australian National University

INDIANA UNIVERSITY
SOUTH BEND