JUNE 18-22, 2023

CVPR VANCOUVER, CANADA

Evading Forensic Classifiers with Attribute-Conditioned Adversarial Faces

Fahad Shamshad, Koushik Srivatsan, Karthik Nandakumar

(WED-PM-392)

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Motivation

- Generative models produce highly realistic synthetic face images, thereby raising security and ethical concerns on digital platforms.

- Face forensic classifiers are developed to defend against these synthetic faces. However, these classifiers are vulnerable to adversarial images.

Limitations of existing works

- Existing methods to generate adversarial images to fool forensic classifiers suffers from the following drawbacks:

1) Contains visible noise patterns which can be detected through human scrutiny.

2) Do not provide face attribute control, that attackers could use to spread false propaganda via social media to specific ethnic or age groups.

3) Effective only in white-box settings.

Goals

- To propose a method that evades forensic classifiers through adversarial faces

  i. with ***specific attributes,***

  *ii. appears **benign to humans**,*

  iii. transferable to ***unknown*** forensic classifiers, and

  iv. provide more control over attributes either via guidance from ***a reference images or a text prompt***.

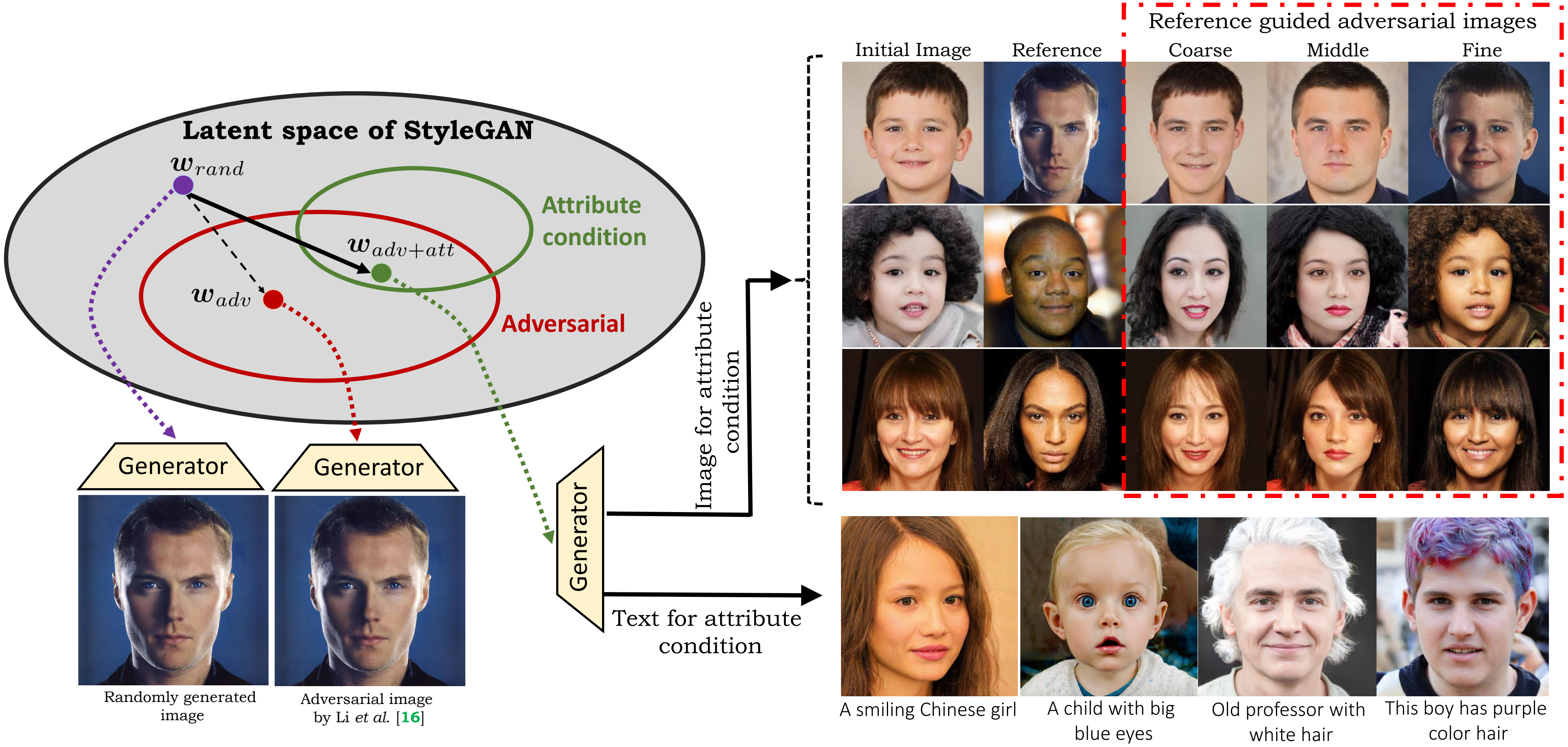
Contributions

- Propose a novel approach to generate adversarial fake faces with a specific set of attributes defined using a reference image or a text prompt.

- Introduced semantic changes that appear benign to humans while being adversarial to deep forensic classifiers.

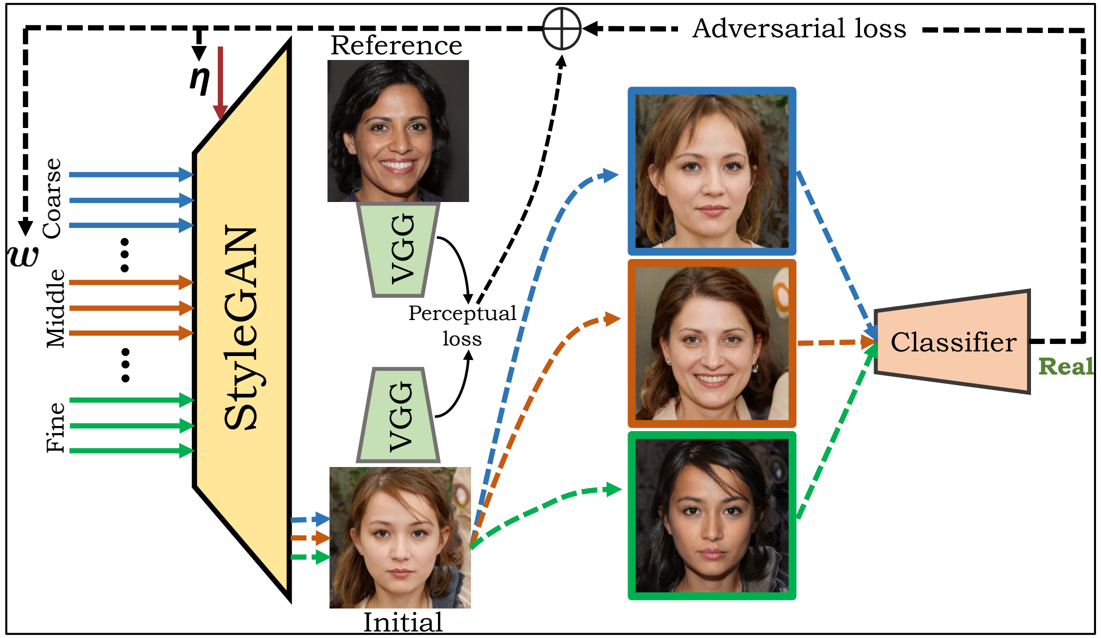
- Meta learning-based optimization strategy to generate adversarial images that are transferable to the unknown forensic classifier models.

An illustration of our attribute-conditioned adversarial face image generation approach.



Method

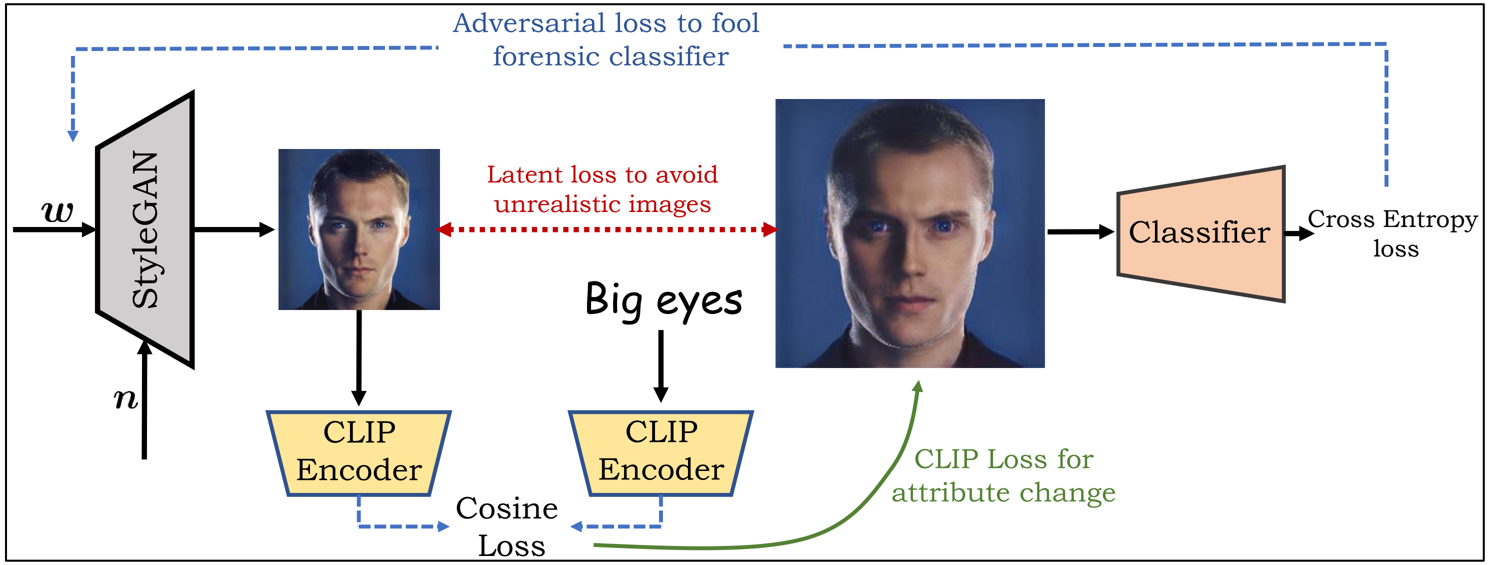
Image as Reference



Adversarially optimize only over the desired attribute-specific layers of the StyleGAN to transfer attributes (pose, expression, or color) to the generated image.

$$(\boldsymbol{\omega}^*, \boldsymbol{\eta}^*) = \arg\min_{\boldsymbol{\omega}, \boldsymbol{\eta}} \|\phi(\mathcal{G}_L(\boldsymbol{\omega}, \boldsymbol{\eta})) - \phi(\mathbf{I}_r)\|_2^2$$
$$+ \lambda_1 \|\boldsymbol{\omega} - \boldsymbol{\omega}_s\|_2^2 + \lambda_2 \mathrm{BCE}(\mathcal{C}(\mathcal{G}_L(\boldsymbol{\omega}, \boldsymbol{\eta})), y = 1)$$

Text as Reference



$$
\begin{aligned}
(\boldsymbol{\omega}^*, \boldsymbol{\eta}^*) = \arg\min_{\boldsymbol{\omega}, \boldsymbol{\eta}} \mathcal{L}_{\text{clip}}(\mathcal{G}_L(\boldsymbol{\omega}, \boldsymbol{\eta}), t) + \lambda_1 \|\boldsymbol{\omega} - \boldsymbol{\omega}_s\|_2^2 \\
+ \lambda_2 \text{BCE}(\mathcal{C}(\mathcal{G}_L(\boldsymbol{\omega}, \boldsymbol{\eta})), y = 1),
\end{aligned}
$$

- For text-guided approach, we leverage the power of rich, joint vision-language representation learned by the **CLIP model**.
- Our optimization scheme aims to modify the latent vector of the StyleGAN under CLIP loss to generate adversarial face images with attributes described by the text prompt.

Meta-Optimization

- We use generic meta-learning-based method to improve transferability to unknown forensic classifiers

- Given a total of T +1 forensic classifiers, we randomly sample T classifiers from them and use T − 1 for meta-train and the remaining model for meta-test. For every iteration, we shuffle and choose different combinations of meta train-test pairs from the set of T classifiers. The latents are first updated to evaluate on the meta-test model, and finally the aggregated losses from the meta-train and meta-test stages are used to optimize the latent for the current iteration.

Results

Method	**Models**					**FID** [13]
	ResNet-18	ResNet-50	VGG-19	DenseNet-121	Wang *et al.* [35]	
Clean accuracy	94%	97%	96%	96%	81%	-
PGD L_{inf} [21]	98%	100%	100%	95%	86%	49.54
FGSM L_{inf} [10]	100%	100%	100%	100%	95%	38.24
Latent (image)	100%	100%	100%	100%	89%	28.31
Noise and latent (image)	100%	100%	100%	100%	100%	26.44
Latent (text)	100%	100%	100%	100%	91%	34.73
Noise and latent (text)	100%	100%	100%	100%	100%	31.92

The attack success rate and FID score of the adversarial images generated by our image-driven and text-guided approaches along with the norm-constrained noise-based methods.

Attribute-conditioned adversarial face images generated via proposed reference image-based approach. All the generated images are misclassified by the forensic classifier

Attribute-conditioned adversarial face images generated via our text-guided method.

Method	ResNet-18	ResNet-50	DenseNet-121	EfficientNet	Xception
Ensemble	11.0	32.0	54.0	46.0	11.0
Meta Learning	**12.0**	**37.0**	**64.0**	**55.0**	**14.0**

ASR in the black box setting. All the models (except the one for which the score is reported) are used during optimization.

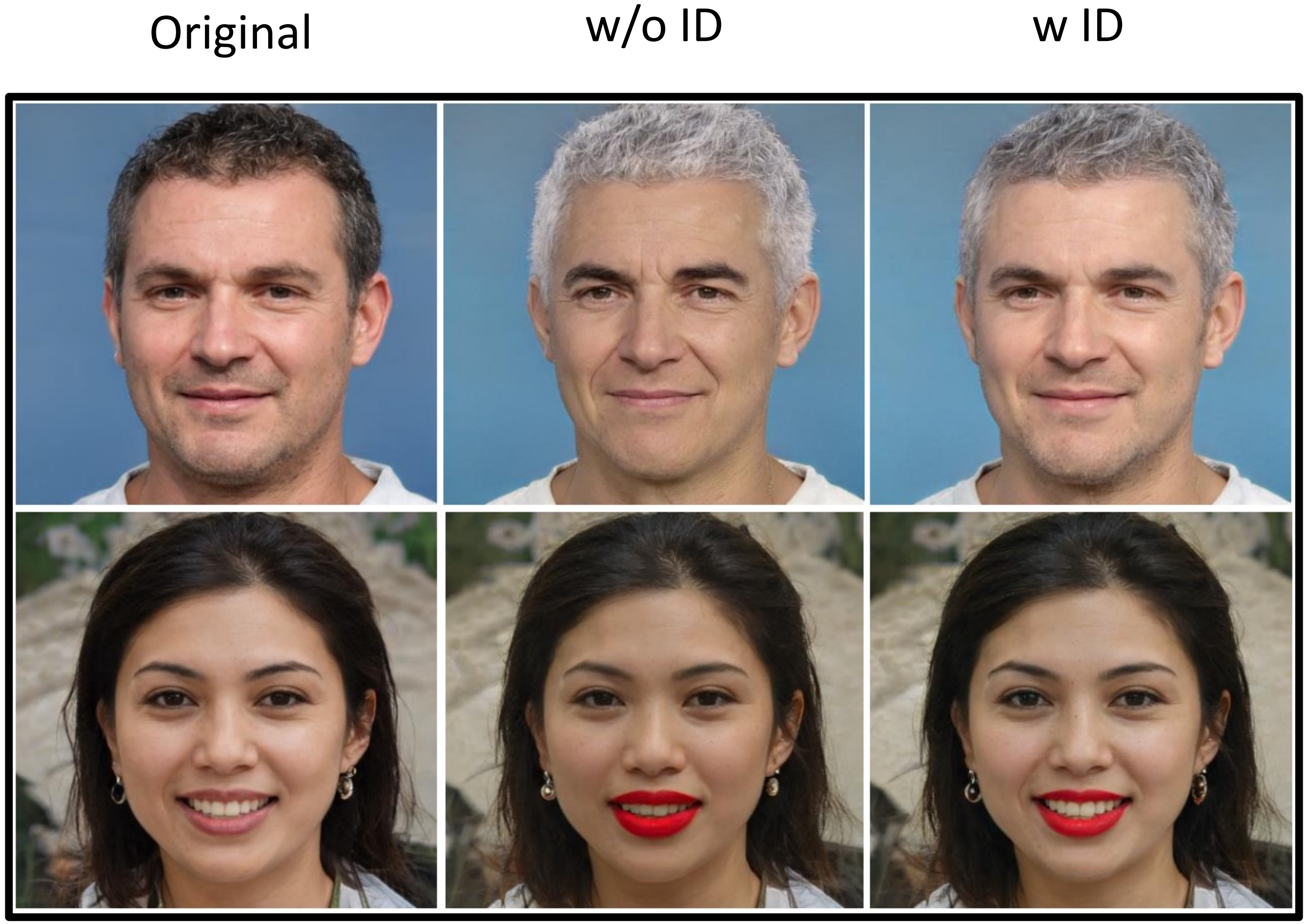
	Time (sec)	*ASR*
Naive	105	100%
Proposed	23	100%

Naive vs proposed image-driven approach.

FGSM [10]	PGD [21]	Proposed (image)	Proposed (text)
1.2%	0%	61%	37.8%

Evaluation of realism of the generated adversarial images by user study.

Additional Results



Effect of using the identity loss.

Attackers can generate diverse images using a specific text prompt (Chinese girl, black skin).

Thank you!