Supplementary Material – Evading Forensic Classifiers with Attribute-Conditioned Adversarial Faces

Fahad Shamshad Koushik Srivatsan Karthik Nandakumar
Mohamed bin Zayed University of AI, UAE
{fahad.shamshad, koushik.srivatsan, karthik.nandakumar}@mbzuai.ac.ae

In this supplementary material, we provide additional implementation details, theoretical explanations and ablation studies to support our contributions in the main paper. In Sec. 1 we provide the details of the text prompts used to evaluate our text-guided approach, followed by the detailed description of the meta-objective function in Sec. 2. In Sec. 3 we provide the quantitative analysis of the effect of changing the values of the hyperparameters defined in our objective functions.

1. Text prompts

To evaluate the proposed text-guided approach, we generate 50 text prompts based on different facial features and ethnicities. Details are given in Table 1. Additional qualitative results of the text-guided approach are provided in Figure 1.

2. Meta Learning based Adversarial Attack

In this section, we provide details of the meta-learning based strategy that we use to improve the black-box transferability of generated attribute-conditioned adversarial image (Sec. 3.3 of the main paper). Our approach consists of two steps: meta-train and meta-test. Below, we provide details of these steps.

Meta-Train: Given T forensic classifiers $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_T$, we randomly sample $T-1$ models to be the meta-train models at each iteration. The meta-train loss for the i^{th} classifier can be defined as,

$$\mathcal{L}_i^{tr}(\boldsymbol{\omega}, \boldsymbol{\eta}) = \text{BCE}(\mathcal{C}_i(\mathcal{G}_L(\boldsymbol{\omega}, \boldsymbol{\eta})), y = 1), \tag{1}$$

where $i \in \{1, ..., T-1\}$, $\mathcal{G}_L(\boldsymbol{\omega}, \boldsymbol{\eta})$ denotes the image generated by StyleGAN that takes latent vector $\boldsymbol{\omega}$, and the noise vector $\boldsymbol{\eta}$ as input. BCE represents the binary cross-entropy loss, and the forensic classifier $\mathcal{C}$ maps an image to an output $y \in \{0, 1\}$, where 0 and 1 represent fake and real classes respectively. For i^{th} model, we update $\boldsymbol{\omega}$ and $\boldsymbol{\eta}$ as,

$$\boldsymbol{\omega}_i^{'} \leftarrow \boldsymbol{\omega} - \alpha_1 \nabla_{\boldsymbol{\omega}} \mathcal{L}_i^{tr}(\boldsymbol{\omega}, \boldsymbol{\eta}) \tag{2}$$

Algorithm 1: Meta Learning based Adversarial Attack

Require:
Input: Generative model $\mathcal{G}_L$; forensic classifier models $\mathcal{C} = \mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{T+1}$; optimization steps K; text prompt t;
Initialization: latent vector $\boldsymbol{\omega}$, and the noise vector $\boldsymbol{\eta}$;
Ensure: $\boldsymbol{\omega}^*, \boldsymbol{\eta}^*$

for each $k \in K$ **do**
Update $\boldsymbol{\omega}$ and $\boldsymbol{\eta}$:
  Calculate $\mathcal{L}_{\text{clip}}$ and $\mathcal{L}_{\text{lat}}$
  Meta-train:
    Randomly select $T-1$ models from $\mathcal{C}$ as meta-train models
    Calculate $\mathcal{L}_i^{tr}$ with Eq.1 for the i^{th} model and then update $\boldsymbol{\omega}_i^{'}$ using Eq.2 and $\boldsymbol{\eta}_i^{'}$ using Eq.3
  Meta-test:
    Use the remaining forensic classifier to be the meta-test model
    Calculate $\mathcal{L}_i^{te}$ with Eq.4
  Meta-optimization:
    With Eq. 5
endfor
return $\boldsymbol{\omega}^*, \boldsymbol{\eta}^*$;

$$\boldsymbol{\eta}_i^{'} \leftarrow \boldsymbol{\eta} - \alpha_1 \nabla_{\boldsymbol{\eta}} \mathcal{L}_i^{tr}(\boldsymbol{\omega}, \boldsymbol{\eta}), \tag{3}$$

where α_1 is the learning rate for the meta update.

Meta-Test: The forensic classifier which is not used during meta-train, $\mathcal{C}_T$, will be the meta-test model. We attack the meta-test model using the updated latents and noise vectors from each of the meta-train models, which can be denoted as,

$$\mathcal{L}_i^{te}(\boldsymbol{\omega}_i^{'}, \boldsymbol{\eta}_i^{'}) = \text{BCE}(\mathcal{C}_T(\mathcal{G}_L(\boldsymbol{\omega}_i^{'}, \boldsymbol{\eta}_i^{'})), y = 1). \tag{4}$$

Meta-Optimization: The latent $\boldsymbol{\omega}$ and noise vectors $\boldsymbol{\eta}$ are finally updated for meta-train and meta-test stages by

Table 1. Text prompts used for the experiments.

	Text prompts		Text prompts
1	Curly long hair	**26**	Old chinese man
2	This man has big nose	**27**	He is from India
3	Bald man with sad face	**28**	She has white hair
4	A girl with mohawk hairstyle	**29**	He has mohawk hairstyle
5	She has big teeth	**30**	This man has long hair
6	He has very big eyes	**31**	She has big nose
7	Woman with heavy makeup	**32**	Old man with blond hair
8	Boy with purple hair	**33**	She is laughing
9	Professor with white hair	**34**	He has curly hair
10	A man with big moustache	**35**	He is bald
11	A smiling Chinese girl	**36**	This girl has black hair
12	This man has black skin	**37**	Smiling chinese man
13	Woman with red hair	**38**	Surprised face
14	Bob-cut hairstyle	**39**	Straight long hair
15	Sad face	**40**	Curly short hair
16	She has blond hair	**41**	Grey hair
17	He has Black hair	**42**	Tanned
18	Baby with big blue eyes	**43**	Pale face
19	She has purple hair	**44**	Wrinkle
20	He has black beard	**45**	Angry face
21	Afro hair	**46**	Bowlcut hairstyle
22	She has big eyes	**47**	Goatee
23	Red lipstick	**48**	Double chin
24	Open mouth	**49**	High cheekbones
25	African face	**50**	Sharp jawline

Table 2. The *attack success rate (ASR)* on the forensic classifier (DenseNet-121) against different values of λ_2, for 100 images generated using the text-driven approach. We set the weightage of latent deviation loss (λ_1) to 0.01.

λ_2	**Attack Success Rate (*ASR*)**
0.05	100
0.0001	54

the following optimization,

$$(\boldsymbol{\omega}^*, \boldsymbol{\eta}^*) = \arg\min_{\boldsymbol{\omega},\boldsymbol{\eta}} \mathcal{L}_{\text{clip}}(\mathcal{G}_L(\boldsymbol{\omega}, \boldsymbol{\eta}), t) + \lambda_1 \|\boldsymbol{\omega} - \boldsymbol{\omega}_s\|_2^2 + \lambda_2 \sum_{i=1}^{T-1} (\mathcal{L}_i^{tr}(\boldsymbol{\omega}, \boldsymbol{\eta}) + \mathcal{L}_i^{te}(\boldsymbol{\omega_i'}, \boldsymbol{\eta_i'})), \quad (5)$$

where, the first term $\mathcal{L}_{\text{clip}}(\cdot, \cdot)$ denotes the cosine distance between the embeddings of the generated image and reference text prompt t, the second term ($\mathcal{L}_{\text{lat}}$) ensures that the manipulated latent codes do not deviate much from the initial latent code $\boldsymbol{\omega}_s$ of the source image and the third term denotes the aggregation of losses from the meta-train and meta-test stages.

The objective function in Eq. (5) can be solved directly using gradient descent method and the final adversarial fake face can be obtained as $\mathbf{I} = \mathcal{G}_L(\boldsymbol{\omega}^*, \boldsymbol{\eta}^*)$. The optimization process is described in Alg. 1.

3. Ablation study

Effect of λ_1 and λ_2: We have conducted experiments by changing the weightage of the latent deviation loss λ_1 and weightage of the adversarial loss λ_2 for our text-based approach. As shown in Figure 2, a large value of λ_1 negatively impacts the attribute faithfulness, and a very small value results in unnatural output images. Similarly, as shown in Table 2, a small value of λ_2 makes it difficult to generate adversarial examples. We observe similar trend for our reference image-based approach.

Number of optimization steps: We evaluate the effect of varying the number of optimization steps for our meta-learning approach as defined in Eq. 5. Figure 3 indicates the increase in average *ASR* with the number of optimization steps.

Number of meta-train models: In Sec. 2, we have

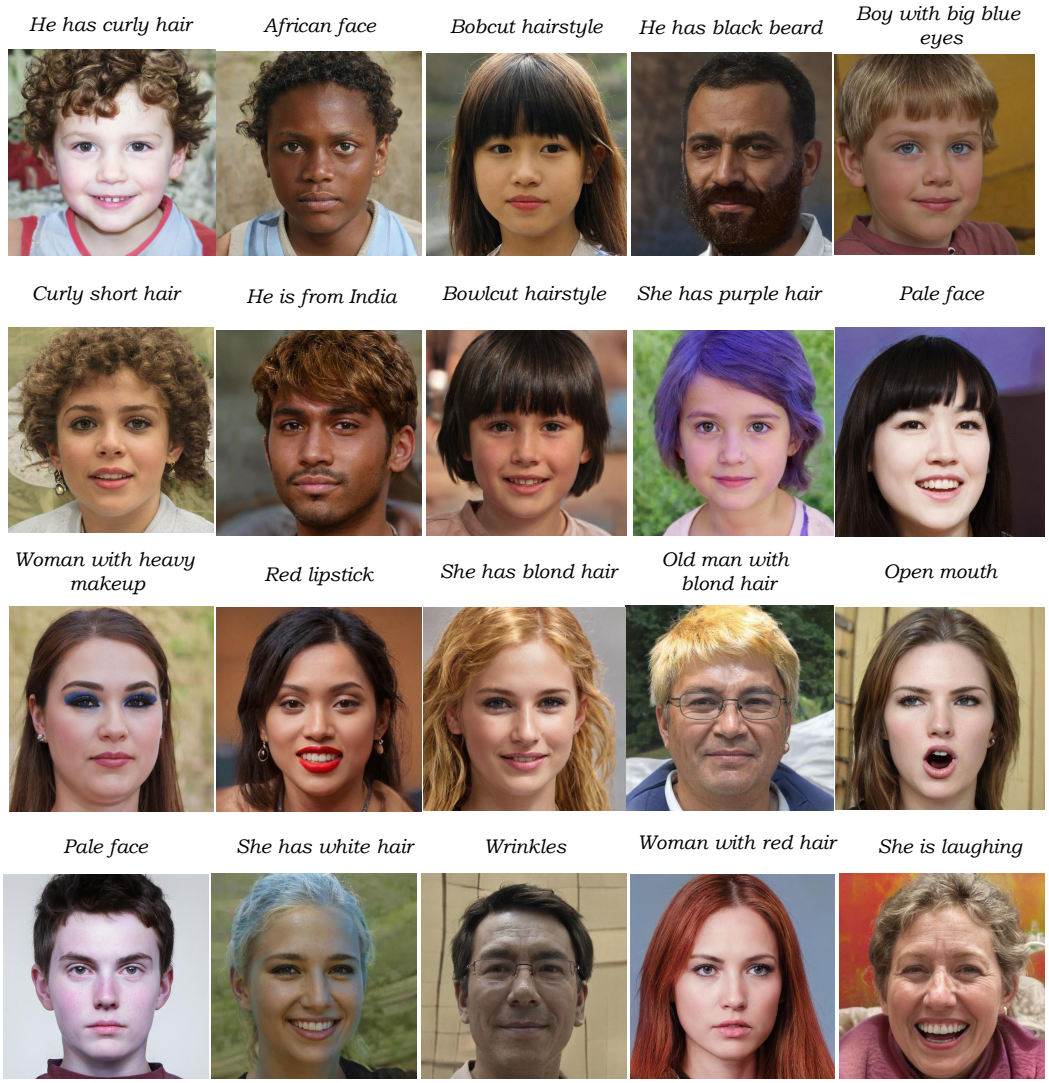

Figure 1. Attribute-conditioned adversarial face images generated via our text-guided method. The driving text prompt is indicated on top of the corresponding image. It can be seen that the generated images contain the attributes provided via the text prompt. Additionally, all these images are misclassified by the forensic model (DenseNet-121).

mentioned that we use only one model in the meta-test stage and the remaining models in the meta-train stage. However, it is also possible to use different combinations of meta-train and meta-test pairs. As shown in Figure 4, the average *ASR* reduces as the number of meta-train model decreases.

4. Societal impact

We introduce attribute-conditioned adversarial attack on human face images to fool deep forensic classifiers. This kind of control over face attributes is essential for attackers to rapidly disseminate false propaganda via social media to specific ethnic or age groups. Since our work focuses on adversarial attacks, in the short run our work can assist various parties with malicious intents to evade forensic classifiers. However, regardless of our work, there is the possibility that such threats will emerge. We believe that in the long run, works such as ours will support further research in developing more robust forensic classifiers that can withstand the kind of attacks we propose, thus negating the short-term risks.

Figure 2. Images generated by changing the weightage term (λ_1) of the latent loss term for the proposed text-guided approach. Text prompt is `Dark black hair`. For the **top row**, we set the value of λ_1 to 0.01. It can be seen that the generated images contain the attributes provided via the text prompt. For the **middle row**, we set the value of λ_1 to 0.1. Giving more weight to λ_1, does not allow the output adversarial images to deviate much from the initial generated images. Therefore, the resulting images do not possess the attributes of the text prompt (also dependent on the initial image). For the **last row**, we set the value of λ_1 to 0.0001. Due to low weightage, the adversarial loss becomes dominant and the output images contain artifacts. Note that all the generated images in the above figure are misclassified by the forensic model (DenseNet-121). We set the weightage of λ_2 to 0.005 .

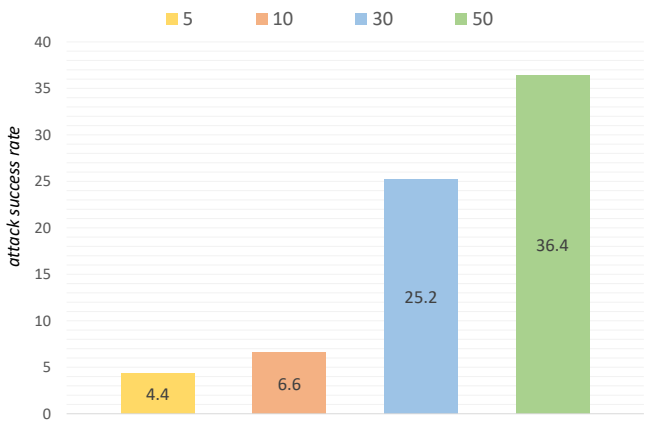

Figure 3. The *ASR* by varying the number of optimization steps in the proposed approach. The results are averaged over five black-box models.

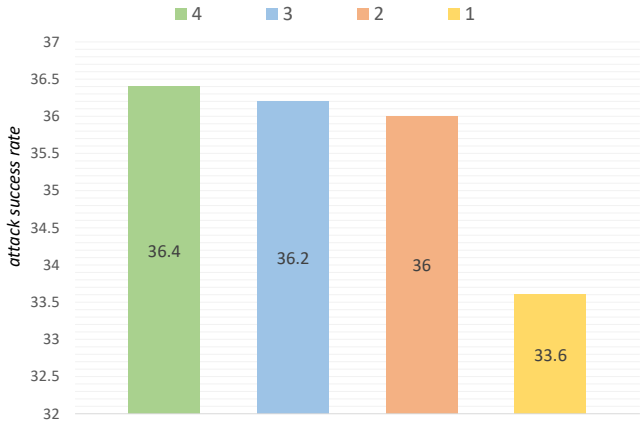

Figure 4. The *ASR* by varying the number of forensic classifiers in the meta-train stage. The results are averaged over five black-box models.