Sibling-Attack: Rethinking Transferable Adversarial Attacks against Face Recognition

Zexin Li[1*] Bangjie Yin[3*] Taiping Yao[3] Junfeng Guo[2] Shouhong Ding[3†] Simin Chen[2] Cong Liu[1†]
[1]University of California, Riverside
[2]The University of Texas at Dallas
[3]Tencent
{zli536, congl}@ucr.edu, {junfeng.guo, simin.chen}@utdallas.edu,
{bangjieyin, taipingyao, ericshding}@tencent.com

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.02359

Abstract

A hard challenge in developing practical face recognition (FR) attacks is due to the black-box nature of the target FR model, i.e., inaccessible gradient and parameter information to attackers. While recent research took an important step towards attacking black-box FR models through leveraging transferability, their performance is still limited, especially against online commercial FR systems that can be pessimistic (e.g., a less than 50% ASR–attack success rate on average). Motivated by this, we present Sibling-Attack*, a new FR attack technique for the first time explores a novel multi-task perspective (i.e., leveraging extra information from multi-correlated tasks to boost attacking transferability). Intuitively,* Sibling-Attack *selects a set of tasks correlated with FR and picks the Attribute Recognition (AR) task as the task used in* Sibling-Attack *based on theoretical and quantitative analysis.* Sibling-Attack *then develops an optimization framework that fuses adversarial gradient information through (1) constraining the cross-task features to be under the same space, (2) a joint-task meta optimization framework that enhances the gradient compatibility among tasks, and (3) a cross-task gradient stabilization method which mitigates the oscillation effect during attacking. Extensive experiments demonstrate that* Sibling-Attack *outperforms state-of-the-art FR attack techniques by a non-trivial margin, boosting ASR by 12.61% and 55.77% on average on state-of-the-art pre-trained FR models and two well-known, widely used commercial FR systems.*

1. Introduction

Deep Neural Networks (DNNs) have demonstrated significant success in various applications, especially for face recognition [14, 53]. Despite these achievements, recent research has revealed that DNN-based face recognition (FR) models may be susceptible to adversarial attacks [2, 23, 59]. In practical attacking scenarios, the victim FR model's parameters are inaccessible to the attackers [4, 19, 43, 52, 61], i.e., the attacker has to perform attacks under black-box settings. One feasible black-box attacking strategy is to craft transferable adversarial examples by attacking a white-box surrogate model. On the face recognition task, recent research (e.g., optimization-based methods [17, 41, 62], model-ensemble training [17, 43] and input data transformations [18, 65, 67]) has shown efficacy on boosting the attacking transferability. Essentially, those methods prevent the adversarial examples from over-fitting to a single model/image by fusing auxiliary gradient information from ensemble models or various sampling/augmenting strategies. However, their performance against online commercial FR

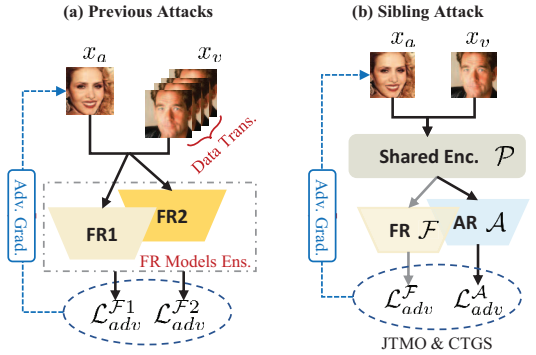

Figure 1. Under the single task, previous attacks **(a)** boost transferability by attacking multiple models or using various sampling or augmentation strategies. Nevertheless, in the proposed Sibling-Attack **(b)**, we adopt the Attribute Recognition (AR) as the auxiliary task to improve the transferability. And we utilize the hard-parameter sharing architecture from [3] as the attacking backbone.

*indicates equal contributions.
†indicates corresponding author.

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.02359

systems can be rather pessimistic (e.g., a less than 50% attack success rate on average as shown in our evaluation).

Motivated by this, we obtain an important insight by understanding such pessimism is that existing methods collect adversarial gradients only from the single task and thus overlook the potential possibilities to further improve transferability, as illustrated in Fig. 1**(a)**. Recent multi-task learning (MTL) methods [3, 42, 58, 74] have indicated that the multi-task or joint-task training among the correlated tasks can learn more robust and general features and thus improve the overall generalizability. Inspired by this, we seek to improve the FR task's attacking transferability within the cross-task scope. To explore the FR attacking transferability under a multi-task setting, there are two challenges: 1) identifying an appropriate auxiliary task as a suitable candidate for FR task when performing multi-task attacks, and 2) how to fully utilize the adversarial information from two tasks thus boosting transferability.

We assume that a face-related task, which can provide relevant but diverse adversarial gradients information to complement the inherently absent adversarial knowledge for the target FR task, could be deemed as a good auxiliary task candidate, named *sibling task*. The empirical observations of previous works [15, 60] have proved that the AR model can learn robust identity features, which can be used to enhance the FR's recognition robustness. Also, in turn, FR features implicitly encode latent facial attribute features. In addition, we conduct quantitative results to show the effectiveness of the AR task. To this end, we leverage a correlated AR task as the sibling task to improve the attacking transferability, i.e., *Sibling-Attack*.

Since big variance exists in the feature and gradient spaces of different tasks [16, 46, 55], direct optimization over FR and AR models will lead to a limited attacking transferability without considering the better gradients fusion and stabilized training strategies. To address the issues, in *Sibling-Attack*, we first adopt the hard-parameter sharing architecture derived from [3] as our backbone attacking framework to constrain them within the same feature space, as shown in Fig. 1**(b)**. Next, we design an alternating joint-task meta optimization (JTMO) algorithm based on the high-level spirit of meta-learning [20, 51, 56] to further improve the gradient compatibility between two tasks. Finally, to mitigate the training oscillation effect, we propose a cross-task gradient stabilization (CTGS) strategy for stabilizing the adversarial example optimization.

Extensive experiments demonstrate that Sibling-Attack outperforms state-of-the-art FR attack techniques by a non-trivial margin, boosting the attack success rate by 12.61% and 55.77% on average on state-of-the-art pre-trained FR models and two well-known, widely used commercial FR systems, Face++ face recognition [48] and Microsoft face API [50]. Notably, Sibling-Attack yields 86.50% and 96.10% ASR on attacking the widely used Face++ commercial face API on two common datasets, while the state-of-the-art only reaches 58.10% and 64.30%, respectively.

We summarize our contributions as: **1)** We propose to generate highly transferable adversarial examples against face recognition by utilizing the adversarial information from the related AR task. **2)** We propose a novel *Sibling-Attack* method which jointly learns the adversarial information from multiple tasks in a more effective manner. **3)** Evidenced by extensive experiments, the ASR of *Sibling-Attack* significantly outperforms current SOTA single-task attacks on the widely-adopted and large-scale FR benchmarks, particularly, several *online commercial FR systems*, which is aligned with our assumptions and analyses.

2. Related Work

2.1. Adversarial Attacks

Adversarial attacks raise significant concern in machine learning due to their potential impact on security and safety-critical applications. [6–8, 10, 17, 24–26, 38, 39, 45, 67, 73] Recently, several approaches have been proposed to enhance the transferability of adversarial attacks by designing underlying optimization algorithms based on the BIM [39] or PGD [45]. For instance, MI-FGSM [17] incorporates momentum to BIM and uses ensemble models to craft adversarial samples. VMI-FGSM [62] alleviates the gradient variance to boost the performance. TAP [75] shows that attacking intermediate feature maps could help to generate more transferable adversarial examples. DI-FGSM [67] proposes a method to increase the diversity of the inputs by randomly altering the input data. Wu. *et al.* [65] makes adversarial examples insensitive to distortions by leveraging a transformation network. Xiong *et al.* [68] focus on reducing stochastic variance to boost ensemble transferable attacking performance. NAA [71] improves the performance of transferable attacks on the feature level by more accurate neuron importance estimations. TAIG [34] boosts transferability by optimizing standard objective functions, exploiting attention maps, and smoothing decision surfaces.

Regarding the transferable digital adversarial attacks against the FR task, Adv-Face [13] employs a GAN-based framework to address the over-fitting problem. DFANet [72] applies dropout layers to boost attacking transferability. On the other hand, a set of work studies transferable physical attacks against FR systems using patch-based methods. Adv-Glasses [57] and Adv-Hat [37] perform physical adversarial attacks by injecting patched hats or eyeglasses. The most recent work [35, 69], generates imperceptible perturbations of specific makeup and facial attributes. Unlike previous work boosting transferability by performing a single task white-box attack, we propose a new framework to craft transferable attacks against

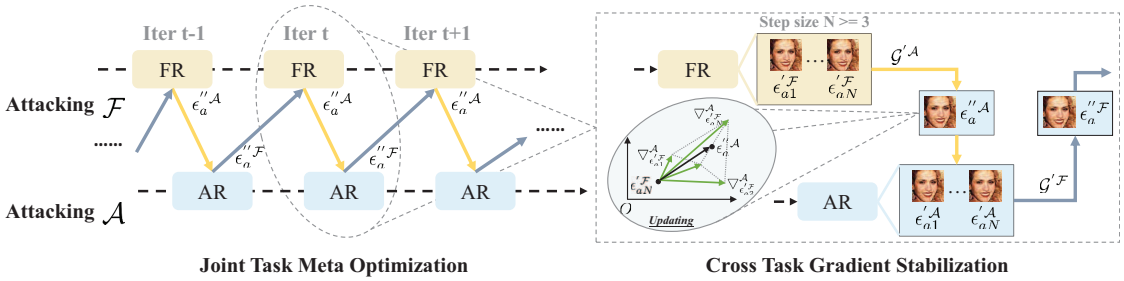

Figure 2. The optimizing process of *Sibling-Attack*. The first row illustrates Joint-Task Meta Optimization (*JTMO*) and the second row exhibits Cross-Task Gradient Stabilization (*CTGS*). JTMO alternatively selects models from different tasks for each iteration (Sec. 3.3). CTGS stabilizes the cross-task gradient via historical information (Sec. 3.4).

the FR model by leveraging the AR task's information.

2.2. Multi-task Learning

Multi-task Learning (MTL) [3, 27, 29, 60, 64] is to learn multiple tasks simultaneously to improve the accuracy of each task compared with single-task learning (STL) [5, 7, 11, 12, 49]. Several existing works have proved the strong correlations between FR and AR tasks. Diniz *et al.* [15] illustrates that the FR model implicitly encodes latent attribute features in the representations, and the hidden layer of the FR model can be used to perform attribute prediction. Hu *et al.* [32] claims that models for the AR task can learn more robust features and thus can be used to improve FR robustness. Taherkhani *et al.* [60] leverage AR models as a soft modality to enhance the performance of FR models. Wang *et al.* [64] utilize a multi-task framework to boost training performance on both FR and AR tasks. Ghamizi et al. [22] and Mao et al. [46] have claimed that multi-task training can learn more adversarial robust features.

Recent concurrent work has studied adversarial attacks against multi-task models. MTA [27] perform white-box attacks adversarial attacks against hard parameter sharing architecture multi-task learning models. UniNet [29] introduces adversarial attacks to better explore the relationship between multi-tasks in an autonomous driving scenario. Nevertheless, there exist significant differences from Sibling-Attack: 1) *Sibling-Attack* focuses on improving black-box attacking transferability rather than maintaining the efficacy of white-box attacks; 2) *Sibling-Attack* proposes JTMO and CTGS optimization strategies to further boost transferability (in Sec. 3). 3) *Sibling-Attack* evaluates transferable attacks against online commercial platforms and significantly improves the performance.

3. Methodology

3.1. Overview

The targeted adversarial attack against FR, i.e., *impersonation attack* [13], spoofs the target FR model to misidentify the attacker as the same identity as the target, which is more challenging and malicious than *dodging attack* [13] in the real world. Therefore, this paper mainly focuses on the impersonation attack as in [69, 72]. The objective of the impersonation attack can be formulated as follows:

$$\min_{\epsilon_a} \mathcal{L}(x_a + \epsilon_a, x_v),\ s.t.\ \|\epsilon_a\|_p \leq \xi \tag{1}$$

where $x_a \in \mathcal{R}^{H\cdot W\cdot C}$ is the attacking face and $x_v \in \mathcal{R}^{H\cdot W\cdot C}$ is the target victim face. The perturbation $\epsilon_a \in [0,1]^{H\cdot W\cdot C}$ to the attacker is constrained by the ℓ_p-norm ($p \in \{0, 2, \infty\}$). In this work, we use ℓ_∞-norm as the metric following [17, 45, 62, 75]. ξ is a small constant to bound ϵ_a. $\mathcal{L}(\cdot)$ denotes the adversarial loss function.

3.2. Sibling-Attack Framework

As shown in Fig. 1**(b)**, we adopt a prevelant hard parameter sharing architecture [1, 3] as the backbone in Sibling-Attack to avoid large feature variance [46]. Our white-box surrogate model shown in Fig. 1**(b)** is denoted as $\mathcal{S}(\mathcal{P};\mathcal{F};\mathcal{A})$, with a sharing-parameter encoder $\mathcal{P}$ as its first component. Then the surrogate model branches off into two sub-networks: an FR branch $\mathcal{F}$, and an AR branch $\mathcal{A}$. Given an attacking image x_a and a target image x_v, our goal is to generate adversarial examples x_{adv} through $\mathcal{S}$ to fool the black-box target FR model $\mathcal{T}$. Specifically, for each x_a and x_v, each branch of $\mathcal{S}$ compute their corresponding output high-level feature vectors $\{f_a^{\mathcal{F}}, f_v^{\mathcal{F}}\}$ through $\mathcal{F}$ and $\{f_a^{\mathcal{A}}, f_v^{\mathcal{A}}\}$ through $\mathcal{A}$, respectively. These features are then used to compute the corresponding adversarial loss for targeted attacks against FR as follows:

$$\mathcal{L}_{adv}^{*} = 1 - cos(f_a^{*}, f_v^{*}) \tag{2}$$

where $* \in \{\mathcal{F}, \mathcal{A}\}$ and we use the cosine value [13, 69, 72] between two feature vectors as the evaluation metric to measure their similarity. Based on that, the main objective of the joint impersonation attack is designed as follows:

$$\min_{\epsilon_a} \lambda_1 \cdot \mathcal{L}_{adv}^{\mathcal{F}} + \lambda_2 \cdot \mathcal{L}_{adv}^{\mathcal{A}},\ s.t.\ \|\epsilon_a\|_p \leq \xi \tag{3}$$

where λ_1 and λ_2 are the trade-off hyper-parameters.

3.3. Joint-Task Meta Optimization

Revisiting the existing meta-learning frameworks, several researchers [21, 51, 56] have proven that alternatively adopting gradients can improve the cross-dataset compatibility of conducting feature learning, thus enhancing generalizability. This fact motivates us to craft transferable adversarial examples by obtaining better gradient compatibility between two tasks. Therefore, we propose a new optimization strategy targeting adversarial scenarios, namely Joint-Task Meta Optimization *(JTMO)*. As shown in Fig. 2, in *JTMO*, we imitate the parameter updating strategy of meta-learning instead of directly calculating weighted average adversarial losses for two tasks.

To generate the adversarial examples, we have to iteratively modify the pixels in x_a by adding a perturbation ϵ_a. For each iteration, we alternately choose one branch from $\mathcal{S}$, and then perform forward- and back-propagation to calculate the gradients from the corresponding adversarial losses, $\mathcal{L}_{adv}^{\mathcal{F}}$ or $\mathcal{L}_{adv}^{\mathcal{A}}$. The order of branch selection won't affect the final performance. For each branch in each iteration, the updated perturbation $\epsilon_a^{'}$ can be computed by:

$$\epsilon_a^{'} \leftarrow \Pi\left\{\epsilon_a - \alpha \cdot sign(\gamma_1 \cdot \bigtriangledown_{\epsilon_a} \mathcal{L}_{adv}^{*}(x_a + \epsilon_a, x_v))\right\} \quad (4)$$

where $\Pi\{\cdot\}$ denotes the projection function ensured by ℓ_∞ constrain. α is learning rate, γ_1 is the updating hyper-parameter, and also $* \in \{\mathcal{F}, \mathcal{A}\}$. Then, we utilize the updated perturbation $\epsilon_a^{'}$ to compute the adversarial losses for the remaining un-chosen branch in the $\mathcal{S}$. Thus, we compute $\mathcal{L}_{adv}^{*'}$ based on $\epsilon_a^{'}$. Finally, we aggregate all the gradient information to update the perturbation as follows:

$$\epsilon_a^{''} \leftarrow \Pi\left\{\epsilon_a^{'} - \alpha \cdot sign(\gamma_2 \cdot \bigtriangledown_{\epsilon_a^{'}} \mathcal{L}_{adv}^{*'}(x_a^{'} + \epsilon_a^{'}, x_v))\right\} \quad (5)$$

where $x_a^{'} = x_a + \epsilon_a$, γ_2 is the updating hyper-parameter and $\epsilon_a^{''}$ is the output of adversarial perturbations for each iteration. Inspired by meta-learning, our optimization strategy first collect gradients alternatively from two branches w.r.t the perturbation parameters, then adopt the gradients to optimize $\epsilon_a^{''}$ alternately between two tasks for every iteration to obtain optimization compatibility.

3.4. Cross-Task Gradient Stabilization

Updating adversarial perturbations across two tasks may inevitably cause a side-effect of oscillation and lead to a sub-optimal solution. This side-effect can be attributed to the fact that the two different tasks have different gradient updating directions [55]. Recent methods of single-task adversarial attacks [17, 62] have claimed that historical gradients and appropriate gradients aggregation could stabilize the optimizing process, thus boosting attacking transferability. Inspired by them, we design a new updating strategy, namely Cross-Task Gradient Stabilizing(*CTGS*), to further improve the attacking transferability of *Sibling-Attack*.

Algorithm 1: The proposed attacking method

Require: Attacking images $x_a \in \mathcal{R}^{H \cdot W \cdot C}$; vcitim images $x_v \in \mathcal{R}^{H \cdot W \cdot C}$; adversarial perturbations $\epsilon_a \in [0,1]^{H \cdot W \cdot C}$; pre-trained multi-task model $\mathcal{S}(\mathcal{P}; \mathcal{F}; \mathcal{A})$; iterations T; updating step size N.
Initialization: Adversarial example parameters ϵ_a;hyperparameters γ_1, γ_2, γ_3; learning rate α.
Ensure: Perturbation parameters ϵ_a^{opt}.
$x_{adv} = x_a$;
for *each* $t \in T$ **do**
  Alternatively select one task branch, such as $\mathcal{F}$;
  Update $\mathcal{E}^{'\mathcal{F}} = \{\emptyset\}$;
  for *each* $i \in N$ **do**
    Calculate $\mathcal{L}_{adv}^{\mathcal{F}}$ on (x_{adv}, x_v) with Eq. 2;
    Obtain $\epsilon_{ai}^{'\mathcal{F}}$ by $\mathcal{L}_{adv}^{\mathcal{F}}$ with Eq. 4;
    Append $\epsilon_{ai}^{'\mathcal{F}}$ to $\mathcal{E}^{'\mathcal{F}}$;
    Update $x_{adv} = x_{adv} + \epsilon_{ai}^{'\mathcal{F}}$;
  Obtain $\mathcal{G}^{'\mathcal{A}} = \left\{\bigtriangledown_{\epsilon_{a1}^{'\mathcal{F}}}^{\mathcal{A}}, ..., \bigtriangledown_{\epsilon_{aN}^{'\mathcal{F}}}^{\mathcal{A}}\right\}$ from another branch $\mathcal{A}$ based on $\mathcal{E}^{'\mathcal{F}}$;
  Update $\epsilon_a^{''\mathcal{A}}$ with Eq. 6;
  Update $x_{adv} = x_{adv} + \epsilon_a^{''\mathcal{A}}$;
$\epsilon_a^{opt} = \epsilon_a^{''\mathcal{A}}$;
return ϵ_a^{opt}

As shown in Fig. 2, at each iteration of the optimizing process, we define an updating step size N for the selected task branch, e.g., $\mathcal{F}$. Then N adversarial perturbations, $\mathcal{E}^{'\mathcal{F}} = \left\{\epsilon_{a1}^{'\mathcal{F}}, ..., \epsilon_{aN}^{'\mathcal{F}}\right\}$, can be crafted iteratively by consecutive steps updating with Eq. 4 based on $\mathcal{F}$. Next, we add the perturbations to the attacking image x_a to generate the adversarial examples and send them into another task branch $\mathcal{A}$ and compute their corresponding gradient maps, $\mathcal{G}^{'\mathcal{A}} = \left\{\bigtriangledown_{\epsilon_{a1}^{'\mathcal{F}}}^{\mathcal{A}}, ..., \bigtriangledown_{\epsilon_{aN}^{'\mathcal{F}}}^{\mathcal{A}}\right\}$. Hence, when updating the $\epsilon_a^{''}$ on the $\mathcal{A}$, we can derive Eq. 5 as:

$$\epsilon_a^{''\mathcal{A}} \leftarrow \Pi\left\{\epsilon_{aN}^{'\mathcal{F}} - \alpha * sign[\gamma_2 * (\bigtriangledown_{\epsilon_{aN}^{'\mathcal{F}}}^{\mathcal{A}} + \gamma_3 \sum_{i=1}^{N-1} \bigtriangledown_{\epsilon_{ai}^{'\mathcal{F}}}^{\mathcal{A}})]\right\} \quad (6)$$

Following this updating procedure, the calculated gradients on $\mathcal{A}$ for the historical adversarial gradients from $\mathcal{F}$ are aggregated for stabilizing the current optimization. γ_3 is a hyper-parameter to balance the training weights. We choose γ_3 as a small number since the historical adversarial gradients merely provide the auxiliary gradients information

Dataset	CelebA-HQ		LFW	
Target Model	IR50	ResNet101	IR50	ResNet101
FR+FR	73.40	76.00	75.80	78.20
FR+FLD	75.20	78.10	52.00	78.60
FR+FP	66.50	85.10	71.80	83.40
FR+AR(Ours)	**93.00**	**93.40**	**97.60**	**96.80**

Table 1. ASR results for black-box impersonation attacks against different task combinations. Best attack performance results are shown in bold. The 2^{nd} place performance is shown in blue.

rather than dominate the main updating direction. Our strategy enhances the optimizing stability and promotes transferability by utilizing the cross-task gradients of the historical $N-1$ adversarial examples from another task branch. The overall procedure of *Sibling-Attack* is shown in Alg. 1.

4. Experiments

4.1. Experimental Setup

Datasets. To evaluate the attacking transferability of the proposed *Sibling-Attack*, we choose two popular face datasets: 1) *CelebA-HQ* [36]: The CelebA-HQ dataset is a high-quality update for the CelebA dataset [44], which consists of 30,000 best-looking facial images. 2) *LFW* [33]: Labeled Faces in the Wild (LFW) is a dataset for face recognition that contains 13,233 images collected on the web of 5,749 different subjects. We randomly sample 1,000 pairs of different-identity faces for each dataset to evaluate *Sibling-Attack*'s attacking performance.

Evaluation Metrics. Following prior works [13,40,69,72], we adopt Attack Success Rate (ASR) for impersonation attack to evaluate *Sibling-Attack*, which is computed through:

$$\text{ASR} = \frac{\text{No. of Comparisons} \geq \tau}{\text{Total No. of Comparisons}} \tag{7}$$

Whether an adversarial attack is successful is defined by the numerator of Eq. 7, which accepts the similarity scores between the adversarial examples and benign examples from the black box model over the corresponding threshold τ.

Baselines. We compare *Sibling-Attack* with ten start-of-the-art adversarial attacks, namely, face-based and transfer-based attacks: 1) *face-based attacks:* Adv-Hat [37], Adv-Glasses [57], Adv-Face [13], Adv-Makeup [69] and GenAP [66]. 2) *transfer-based attacks:* PGD [45], TAP [75], MI-FGSM [17], VMI-FGSM [62].

Target Model. Similar to the evaluation in prior works [69], we choose a mix of the various offline and online commercial FR models to evaluate the transferability of the adversarial examples generated by *Sibling-Attack*. Specifically, we choose: 1) *Offline models:* five famous face recognition models: IR152 [14], IRSE50 [14], FaceNet [53], IR50 [14], ResNet101 [31]. 2) *Online models:* two widely used online commercial face recognition systems: Face++ [48] and Microsoft [50]. For the offline FR models, we use IR152, IRSE50, and FaceNet as white-box models to generate adversarial examples and evaluate attacking transferability on the other models. All the thresholds of offline models are obtained from the images in the LFW dataset [13]. We set τ to (0.277, 0.200) following [45, 69, 70] for (IR50, ResNet101). For the online FR models, as per the suggestions of platforms, we set τ as the cosine similarity score at 0.001 FPR (False Positive Rate) level for Face++. For Microsoft, We use the reported query results as the number of successful attacks since Microsoft does not offer the cosine similarity score for different FPR levels and only gives a cosine similarity score and decision result for each query.

AR Model. For AR models, we use IR152 [14] and Mobileface [9] as the backbone networks and train them on MS-Celeb-1M [28], and CelebA-HQ [36], to guarantee their performance on the AR task. We include the detailed training scheme in the supplementary files due to the page limits.

Implementation Details. In *Sibling-Attack*, the structure of the white-box surrogate model is IR152 for both the FR task and the AR task. Following the experimental configuration of previous work [66], we set ξ to 40/255 as the ℓ_∞ bound as [17,30,45,62,75] for ours and baselines. Meanwhile, the step size α is set to 2/255 while the iteration number T is set to 200 to ensure attack efficacy, and updating step size N is 4. Moreover, we initialize $(\gamma_1, \gamma_2, \gamma_3)$ as (0.1, 0.9, 0.01). All competitors strictly adopt their original setting. Since our method performs attacks across two different models, FR and AR, and existing works [17, 43] have evidenced the merits of ensemble attacking, we attack two FR models for other competitors to ensure comparison fairness.

4.2. Why Select the AR Task?

Theoretically, we have presented the high correlations between FR and AR in the earlier sections. Furthermore, we empirically explore the effectiveness of the AR task by quantitative analysis. Firstly, we compare the transferable ASRs results against various face-related task combinations in Tab. 1, where FR denotes the *Face Recognition* task, FLD indicates the *Face Landmark Detection* task, FP means the *Face Parsing* task, and AR denotes the *Attribute Recognition* task. And then, all the combinations follow the basic Hard Parameter Sharing architecture to construct the joint-task attacking framework. Their attacking losses are in the same form as $\mathcal{L}^*_{adv}$ in Eq. 2. The ASR results demonstrate that the FR+AR outperforms all the competitors, which quantitatively proves that leveraging the AR task as a sibling task can craft more effective adversarial attacks. And the under-performance of FR+FLD and FR+FP combinations also indicate that not all face-related tasks can contribute to FR attacking transferability. In turn, it confirms

Methods	Dataset	CelebA-HQ							
	Source Model	IR152+FaceNet				IR152+IRSE50			
	Target Model	Offline Model		Online Model		Offline Model		Online Model	
		IR50	ResNet101	Face++	Microsoft	IR50	ResNet101	Face++	Microsoft
Face-based	Adv-Hat [37]	1.50	6.50	1.00	0.00	3.80	8.70	0.90	0.00
	Adv-Glasses [57]	0.60	8.50	3.40	0.00	5.90	9.70	4.20	0.10
	Adv-Face [13]	58.80	64.60	54.90	8.70	68.00	71.40	48.00	8.70
	Adv-Makeup [69]	8.30	21.20	5.30	0.00	13.00	26.00	4.90	0.10
	GenAP [66]	52.80	49.10	54.40	6.40	47.10	48.40	47.20	5.80
Transfer-based	PGD [45]	73.40	76.00	37.20	13.00	92.00	90.80	58.10	28.70
	TAP [75]	72.80	76.20	42.90	20.40	88.30	87.60	52.90	28.90
	MI-FGSM [17]	66.60	73.30	36.10	14.80	86.20	90.10	57.80	28.90
	VMI-FGSM [62]	78.20	83.20	35.70	7.20	80.80	82.90	38.70	9.70
Ours	*Sibling-Attack*	**94.10**	**93.70**	**86.50**	**34.50**	**94.10**	**93.70**	**86.50**	**34.50**
		15.90 ↑	10.50 ↑	31.60 ↑	14.10 ↑	2.10 ↑	2.90 ↑	28.40 ↑	5.60 ↑

Table 2. ASR results of black-box impersonation attack over CelebA-HQ dataset. Two offline models and two online commercial FR systems (Face++ and Microsoft) are used to evaluate attacking transferability. Our method uses IR152 FR and IR152 AR for white-box training, while other methods for comparisons are trained using two different FR models. The best-attacking performance results are shown in bold. The 2^{nd} place performance is shown in blue. The last row shows the promotion between best results vs. 2^{nd} results.

Methods	Dataset	LFW							
	Source Model	IR152+FaceNet				IR152+IRSE50			
	Target Model	Offline Model		Online Model		Offline Model		Online Model	
		IR50	ResNet101	Face++	Microsoft	IR50	ResNet101	Face++	Microsoft
Face-based	Adv-Hat [37]	1.80	9.30	1.80	0.10	5.00	13.40	2.20	0.10
	Adv-Glasses [57]	0.80	5.00	3.70	0.00	1.90	4.90	4.70	0.00
	Adv-Face [13]	13.80	29.70	30.70	0.40	13.80	24.80	19.00	0.40
	Adv-Makeup [69]	2.40	9.20	5.30	0.20	4.70	12.60	5.50	0.30
	GenAP [66]	4.20	13.60	15.20	0.30	4.30	14.50	13.90	0.50
Transfer-based	PGD [45]	75.80	78.20	46.70	19.10	89.30	89.70	60.40	36.50
	TAP [75]	76.90	81.00	54.10	28.60	89.60	89.60	64.30	45.60
	MI-FGSM [17]	68.40	71.00	41.90	21.10	92.20	86.30	60.10	38.80
	VMI-FGSM [62]	76.80	80.80	41.50	10.90	76.40	79.30	40.80	11.90
Ours	*Sibling-Attack*	**98.70**	**98.60**	**96.10**	**59.30**	**98.70**	**98.60**	**96.10**	**59.30**
		21.80 ↑	17.60 ↑	42.00 ↑	30.70 ↑	6.50 ↑	8.90 ↑	31.80 ↑	13.70 ↑

Table 3. ASR results of black-box impersonation attack over LFW dataset. The settings are following Tab. 2.

the necessity of selecting appropriate face-related tasks as the attacking candidates for the FR task.

4.3. Experimental Results

Comparison with face-based methods. From Tab. 2 and 3, we observe that the patch-based methods have weak transferability on most target models as they are designed and tuned for physical attacks with small attacking areas. The results show that the adversarial examples attacking the entire face, i.e., Adv-Face, have the best transferability compared to all the other face-based methods. However, *Sibling-Attack* can still significantly outperform Adv-Face.

Comparison with transfer-based methods. We then compare *Sibling-Attack* with four transfer-based attack methods (designed to generate strongly transferable adversarial examples). As observed from the results of CelebA-HQ in Tab. 2, *Sibling-Attack* dominates all the transfer-based methods across various settings and evaluated models. Specifically, under the setting that uses IR152+FaceNet as white-box models, *Sibling-Attack* outperforms the best results of transfer-based methods under offline models by 15.90% on IR50 and 10.50% on ResNet101. Meanwhile, *Sibling-Attack* outperforms the best results of other competitors under online models by 31.60% on Face++ and 14.10% on Microsoft. Similarly, Tab. 2 and 3 also show our superior performance. On average, *Sibling-Attack* improves the state-of-the-art ASRs by 12.61% and 55.77% for offline pre-trained and online commercial models.

Methods	Dataset			LFW			
	Source Model			Offline Model		Online Model	
	IR152	FaceNet	IRSE50	IR50	ResNet101	Face++	Microsoft
Single Model	✓	-	-	76.50	79.30	43.40	13.10
	-	✓	-	1.30	5.10	4.90	0.20
	-	-	✓	63.40	76.80	56.50	14.20
Ensemble	✓	✓	-	75.80	78.20	46.70	19.10
	✓	-	✓	89.30	89.70	60.40	36.50
	-	✓	✓	65.80	77.90	59.20	16.80
Ours	Basic framework			80.90	92.20	69.80	37.20
	+ Hard P.S.			97.60	96.80	77.40	45.40
	+ JTMO			98.30	98.40	95.50	51.20
	+ CTGS			**98.70**	**98.60**	**96.10**	**59.30**

Table 4. Comparisons of ASR results of impersonation attack over LFW dataset. The ensemble represents the ensemble-training-based method. The 2^{nd} place results are shown in blue.

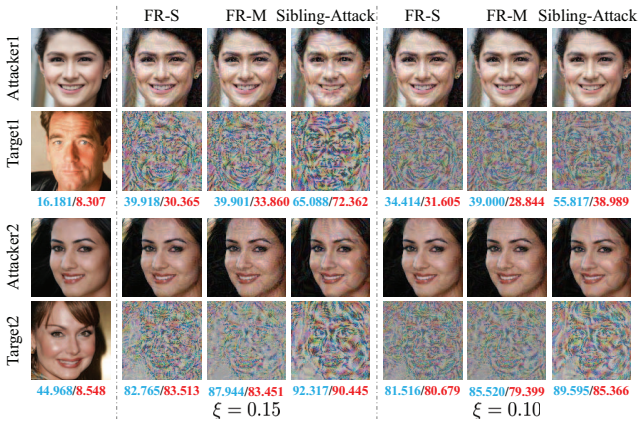

Figure 3. Visualization of our adversarial perturbations comparing with attacks only against FR models. Each column shows the adversarial example and its post-processed perturbations. Query results from Face++ and Microsoft are shown in blue and red.

4.4. Ablation Study

We study the impact of the different components of *Sibling-Attack*. Specifically, *Sibling-Attack* consists of the following components: (a) Hard Parameter Sharing (denoted as "Hard P.S."), (b) Joint-Task Meta Optimization (JTMO), and (c) Cross-Task Gradient Stabilizing (CTGS). As shown in Tab. 4, we investigate the performance of *Sibling-Attack* with different incorporated components. The line, "*Basic framework*", is to directly average the adversarial losses for FR and AR models without optimization strategies. Its ASRs are competitive under single/ensemble model (single-task) settings. Empirically, this experimental result strongly supports the effectiveness of our idea that using the information from AR tasks can help boost attacking transferability against the FR task. Besides, we can observe that the ASRs gradually increase with adding each proposed component and significantly outperform other single/ensemble training-based competitors. Specifically, for two online models, Hard P.S. architecture boosts ASRs by 7.60% and 8.20%, JTMO improves 18.10% and 5.80% compared with the Hard P.S. architecture. CTGS further achieves improvements of 0.60% and 8.10% compared with the Hard P.S. with JTMO. The results demonstrate the effectiveness of each proposed component in *Sibling-Attack*.

4.5. Visualization and Analysis

Visualization of adversarial perturbations. In Fig. 3, we visualize the generated adversarial examples/perturbations from FR-S (against a single FR model, IR152), FR-M (against two ensemble FR models, IR152 and FaceNet), and *Sibling-Attack* for two pairs of target and attacker images from CelebA-HQ. For fair comparisons, we select a cross-gender and a same-gender pair. The first column presents the legitimate examples with query results, and the following columns present the adversarial examples/perturbations under two different L_∞ bounds (ξ=0.10,0.15) with the query results from Face++ and Microsoft, respectively. Specifically, we post-process the perturbations to make them more perceptible. In detail, we multiply all the perturbation values by 5 then truncate the values less than $\xi/3$, then project each perturbation value into $[0, 255]$ for better visualization. We can discern a salient shape of a face and some facial components in the adversarial perturbations generated by *Sibling-Attack*, which are different from the perturbations generated by FR-S and FR.

Visualization of black-box adversarial gradient responses. We further explore why adversarial examples generated by *Sibling-Attack* exhibit more attacking transferability by employing Grad-CAM [54], as shown in Fig. 4. For each row, we visualize the gradient responses. Specifically, FR-B (Black) denotes the black-box scenario, ensemble attacking IR152 and FaceNet using PGD and visualizing Grad-CAM on IRSE50. FR-W (White) denotes the white-box scenario, ensemble attacking IRSE50 via PGD and visualizing Grad-CAM on IRSE50. Notably, the gradient responses in FR-W serve as the ground truth for measuring the attacking transferability of each approach. Specifically, the more visual similarity in gradient responses between the evaluated approach and ground truth implies stronger transferability. We can observe that gradient responses in FR-B seem either (1) to pay more attention to the background or (2) overfit to some local facial regions. In contrast, gradient responses from *Sibling-Attack* and the target model both focus more on the similar key facial regions, which interprets the stronger transferability of *Sibling-Attack*.

Visually-indistinguishable analysis. In addition to the efficacy, we also analyze the visual indistinguishability of our crafted adversarial samples. We use *Structural Similarity* (SSIM) [63] and the *Mean Square Error* (MSE) [47] between basic examples and corresponding adversarial examples as metrics. As shown in Table. 5, we compare

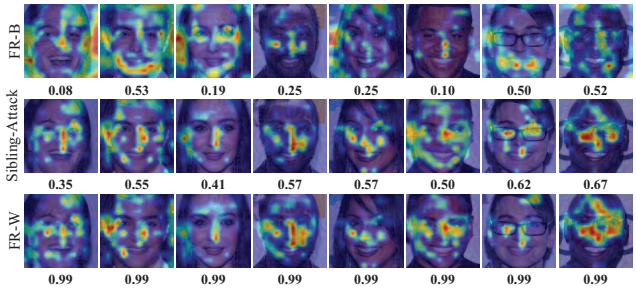

Figure 4. Visualization using Grad-CAM [54] produces attention maps on an offline FR model (IRSE50). We display the similarity score between the attacker and the target face on the FR model under each picture. Best viewed in color.

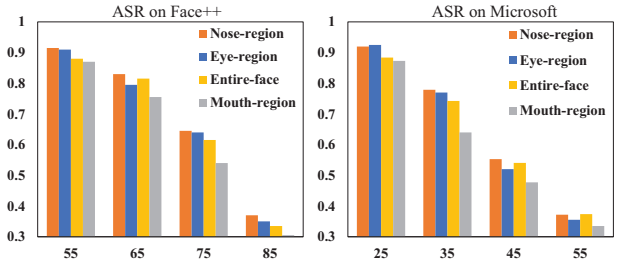

Figure 5. ASR on Face++ and Microsoft of *Sibling-Attack* with AR models trained by different facial attribute groups. The x-axis represents the similarity score. The y-axis represents the ASR under the corresponding similarity score.

SSIM and MSE with the other invisible attacking methods on LFW. The results demonstrate that the SSIM and MSE of *Sibling-Attack* are competitive with other methods. Last but not least, *Sibling-Attack* can achieve a much better-attacking transferability against black-box FR models.

Transferability analysis of different facial attributes. Our experimental results for *Sibling-Attack* have already shown that exploiting an auxiliary AR model can help generate strongly transferable adversarial examples against FR models. This section further explores which facial attributes can bring more transferability to the FR task. Specifically, we divide the 18 facial attributes used for training our IR152 AR model into four non-overlapping groups by position (eye-region, nose-region, mouth-region, other-region). As Fig. 5 shows, mouth-region facial attributes bring weaker transferability to the FR task than attributes in other regions, which is consistent with some existing works [15, 66] on face recognition and face-based attacks.

Transferability analysis against AR tasks. We also explore whether attackers can adopt adversarial information from the FR model to improve the transferability against the black-box target AR model by *Sibling-Attack*. As shown in Tab. 6, we compare the overall attributes prediction difference across 1000 image pairs for four attribute groups. The

Dataset	LFW			
Source Model	IR152+FaceNet		IR152+IRSE50	
Metrics	SSIM	MSE	SSIM	MSE
PGD [45]	0.619	**175.915**	0.594	193.801
TAP [75]	0.613	181.279	0.591	196.942
MI-FGSM [17]	0.473	343.227	0.463	350.162
VMI-FGSM [62]	0.588	200.418	0.574	215.346
Sibling-Attack	**0.626**	187.491	**0.626**	**187.491**

Table 5. SSIM and MSE scores of our methods and other competitors. The best results are shown in bold. The 2^{nd} place performance is shown in blue.

Group	Eye-region	Nose-region	Mouth-region	Other-region
Baseline AR	148.85	223.97	184.77	201.79
Ours	**162.41**	**241.29**	**195.46**	**214.02**

Table 6. Comparisons of the overall prediction difference between *Sibling-Attack* and a baseline white-box AR for four attribute groups after attacking a black-box AR model.

overall attributes prediction changes can be computed as:

$$\text{Overall Pred. Diff.} = \sum_{s}^{S} \|\mathcal{A}_B(x^s_{adv}) - \mathcal{A}_B(x^s)\|_1 \quad (8)$$

where $\mathcal{A}_B(\cdot)$ denotes the target AR model, which outputs the predicting score for each attribute, and $S = 1000$, x^s_{adv} is the adversarial example crafted by attacking our model or the baseline white-box AR model. The results indicate that *Sibling-Attack* can conduct more prediction difference than the competitor, which supports that our method can also boost the attacking transferability against the AR model.

5. Conclusion.

The proposed *Sibling-Attack* firstly leverages a highly FR-related task AR as the sibling task to generate strongly transferable adversarial attacks against FR tasks under the black-box setting. It mainly focuses on digital scenarios, but it is equally essential for face recognition security as to the physical attacks since it can reveal more threatening adversarial risks. Besides, the proposed method may be used maliciously to hazard the security of existing FR models in real life, the adversarial training and de-noise strategies can mitigate the negative impacts. Extensive experiments demonstrate the superior transferability of *Sibling-Attack* on various offline and online commercial FR models. In the future, we also intend to extend the proposed idea to other computer vision and biometrics tasks besides FR.

Acknowledgments

This work was supported by NSF CNS 2135625, CPS 2038727, CNS Career 1750263, and a Darpa Shell grant.

References

[1] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997. 3

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1

[3] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 1, 2, 3

[4] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop On Artificial Intelligence and Security*, pages 15–26, 2017. 1

[5] Simin Chen, Soroush Bateni, Sampath Grandhi, Xiaodi Li, Cong Liu, and Wei Yang. Denas: automated rule generation by knowledge extraction from neural networks. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 813–825, 2020. 3

[6] Simin Chen, Mirazul Haque, Cong Liu, and Wei Yang. Deepperform: An efficient approach for performance testing of resource-constrained neural networks. In *37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13, 2022. 2

[7] Simin Chen, Hamed Khanpour, Cong Liu, and Wei Yang. Learn to reverse dnns from AI programs automatically. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 666–672. ijcai.org, 2022. 2, 3

[8] Simin Chen, Cong Liu, Mirazul Haque, Zihe Song, and Wei Yang. Nmtsloth: understanding and testing efficiency degradation of neural machine translation systems. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1148–1160, 2022. 2

[9] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Proceedings of the Chinese Conference on Biometric Recognition (CCBR)*, pages 428–438, 2018. 5

[10] Simin Chen, Zihe Song, Mirazul Haque, Cong Liu, and Wei Yang. Nicgslowdown: Evaluating the efficiency robustness of neural image caption generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15365–15374, 2022. 2

[11] Yiming Chen, Yan Zhang, Bin Wang, Zuozhu Liu, and Haizhou Li. Generate, discriminate and contrast: A semi-supervised sentence representation learning framework. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8150–8161, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 3

[12] Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 3

[13] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 2, 3, 5, 6

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 1, 5

[15] Matheus Alves Diniz and William Robson Schwartz. Face attributes as cues for deep face recognition understanding. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 307–313. IEEE, 2020. 2, 3, 8

[16] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and HT Kung. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19217–19227, 2022. 2

[17] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018. 1, 2, 3, 4, 5, 6, 8

[18] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4312–4321, 2019. 1

[19] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7714–7722, 2019. 1

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017. 2

[21] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017. 4

[22] Salah Ghamizi, Maxime Cordy, Mike Papadakis, and Yves Le Traon. Adversarial robustness in multi-task learning: Promises and illusions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 697–705, 2022. 3

[23] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[24] Junfeng Guo, Ang Li, and Cong Liu. AEVA: Black-box backdoor detection using adversarial extreme value analy-

sis. In *International Conference on Learning Representations*, 2022. 2

[25] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. SCALE-UP: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[26] Junfeng Guo and Cong Liu. Practical poisoning attacks on neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 142–158. Springer, 2020. 2

[27] Pengxin Guo, Yuancheng Xu, Baijiong Lin, and Yu Zhang. Multi-task adversarial attack. *arXiv preprint arXiv:2011.09824*, 2020. 3

[28] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102. Springer, 2016. 5

[29] Naresh Kumar Gurulingan, Elahe Arani, and Bahram Zonooz. Uninet: A unified scene understanding network and exploring multi-task relationships through the lens of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2239–2248, 2021. 3

[30] Bangyan He, Jian Liu, Yiming Li, Siyuan Liang, Jingzhi Li, Xiaojun Jia, and Xiaochun Cao. Generating transferable 3d adversarial point cloud via random perturbation factorization. In *AAAI*, 2023. 5

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[32] Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S Mukherjee, Timothy M Hospedales, Neil M Robertson, and Yongxin Yang. Attribute-enhanced face recognition with neural tensor fusion networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3744–3753, 2017. 3

[33] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 5

[34] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. In *The Tenth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2022. 2

[35] Shuai Jia, Bangjie Yin, Taiping Yao, Shouhong Ding, Chunhua Shen, Xiaokang Yang, and Chao Ma. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems (NIPS)*, 2022. 2

[36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. 5

[37] S. Komkov and A. Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society. 2, 5, 6

[38] Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14254–14263, 2020. 2

[39] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 2

[40] Yiming Li, Baoyuan Wu, Yan Feng, Yanbo Fan, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Semi-supervised robust training with generalized perturbed neighborhood. *Pattern Recognition*, 124:108472, 2022. 5

[41] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2020. 1

[42] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880, 2019. 2

[43] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of 5th International Conference on Learning Representations (ICLR)*, 2017. 1, 5

[44] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference On Computer Vision (ICCV)*, pages 3730–3738, 2015. 5

[45] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018. 2, 3, 5, 6, 8

[46] Chengzhi Mao, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. Multitask learning strengthens adversarial robustness. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 (ECCV)*, pages 158–174. Springer, 2020. 2, 3

[47] Hans Marmolin. Subjective mse measures. *IEEE transactions on systems, man, and cybernetics*, 16(3):486–489, 1986. 7

[48] MEGVII. Online face verification. https://www.faceplusplus.com.cn/, 2021. 2, 5

[49] Fei Miao, Sihong He, Lynn Pepin, Shuo Han, Abdeltawab Hendawi, Mohamed E Khalefa, John A Stankovic, and George Pappas. Data-driven distributionally robust optimization for vehicle balancing of mobility-on-demand systems. *ACM Transactions on Cyber-Physical Systems*, 5(2):1–27, 2021. 3

[50] Microsoft. Online face verification. https://azure.microsoft.com/, 2021. 2, 5

[51] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018. 2, 4

[52] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS)*, pages 506–519, 2017. 1

[53] Florian Schroff, Kalenichenko Dmitry, and Philbin James. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 5

[54] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 618–626, 2017. 7, 8

[55] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 2, 4

[56] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11974–11981, 2020. 2, 4

[57] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security (CCS)*, pages 1528–1540, 2016. 2, 5, 6

[58] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning (ICML)*, pages 9120–9132. PMLR, 2020. 2

[59] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[60] Fariborz Taherkhani, Nasser M Nasrabadi, and Jeremy Dawson. A deep face identification network enhanced by facial attributes prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPR)*, pages 553–560, 2018. 2, 3

[61] Wenxuan Wang, Bangjie Yin, Taiping Yao, Li Zhang, Yanwei Fu, Shouhong Ding, Jilin Li, Feiyue Huang, and Xiangyang Xue. Delving into data: Effectively substitute training for black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4761–4770, 2021. 1

[62] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1924–1933, 2021. 1, 2, 3, 4, 5, 6, 8

[63] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002. 7

[64] Zhanxiong Wang, Keke He, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR)*, pages 365–374, 2017. 3

[65] Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9024–9033, 2021. 1, 2

[66] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11845–11854, 2021. 5, 6, 8

[67] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2019. 1, 2

[68] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14983–14992, 2022. 2

[69] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Advmakeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1252–1258. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 2, 3, 5, 6

[70] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482. PMLR, 2019. 5

[71] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14993–15002, 2022. 2

[72] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security (TIFS)*, 16:1452–1466, 2020. 2, 3, 5

[73] Husheng Zhou, Wei Li, Zelun Kong, Junfeng Guo, Yuqun Zhang, Bei Yu, Lingming Zhang, and Cong Liu. Deepbillboard: Systematic physical-world testing of autonomous driving systems. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pages 347–358, 2020. 2

[74] Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4514–4523, 2020. 2

[75] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018. 2, 3, 5, 6, 8