Sibling-Attack: Rethinking Transferable Adversarial Attacks Against Face Recognition

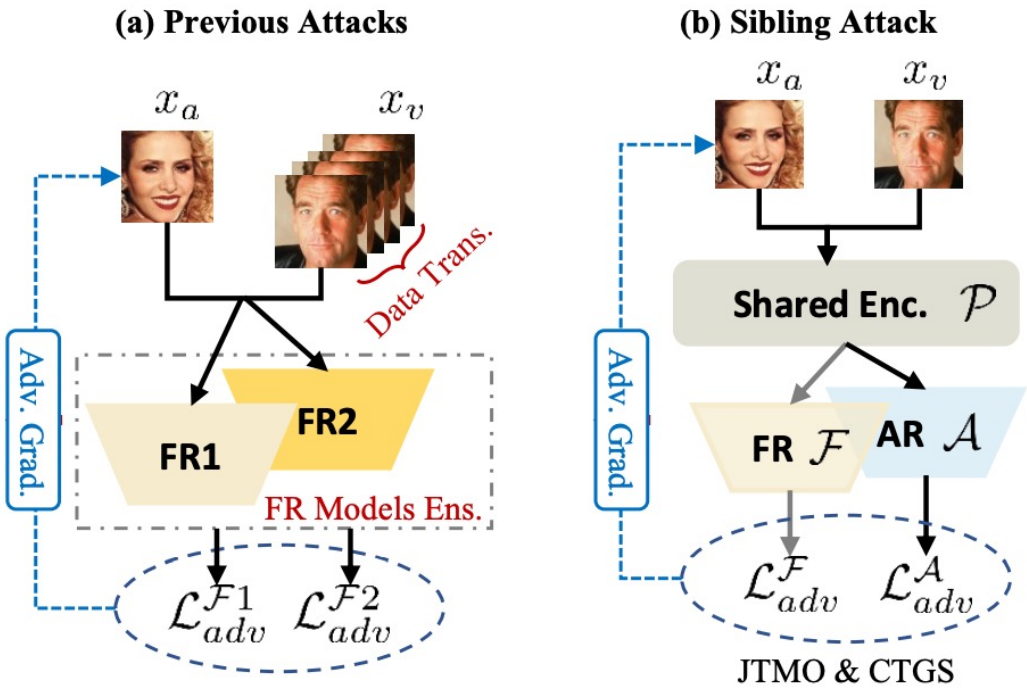
Zexin Li[1], Bangjie Yin[3], Taiping Yao[3], Junfeng Guo[2], Shouhong Ding[3], Simin Chen[2], and Cong Liu[1]
[1]University of California, Riverside, [2]The University of Texas at Dallas, [3]Youtu Lab, Tencent

Poster ID: THU-PM-385

UC RIVERSIDE

Introduction

- Transferable adversarial attack against face recognition (FR) task.
- Leverage adversarial information from multiple tasks.



UC RIVERSIDE

Auxiliary Task Selection

- Theoretical analysis conducted in previous works supports FR and AR are highly correlated tasks.
- Empirical analysis demonstrates that AR exhibits best attacking transferability performance for intuitive multi-task attack.

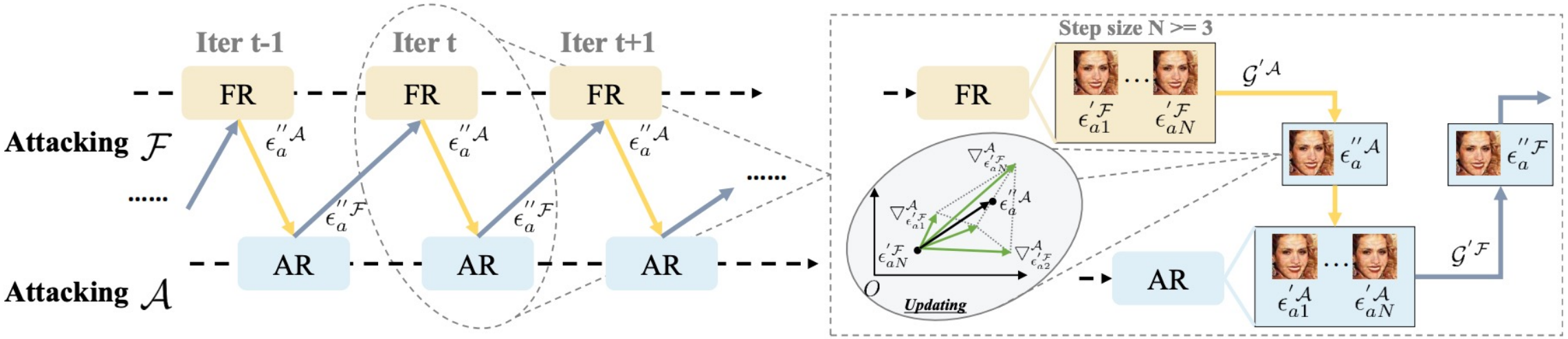
Dataset	CelebA-HQ		LFW	
Target Model	IR50	ResNet101	IR50	ResNet101
FR+FR	73.40	76.00	75.80	78.20
FR+FLD	75.20	78.10	52.00	78.60
FR+FP	66.50	85.10	71.80	83.40
FR+AR(Ours)	**93.00**	**93.40**	**97.60**	**96.80**

*FR: face recognition; AR: facial attribute recognition; FP: face parsing; FLD: face landmark detection.

UC RIVERSIDE

Optimization Framework

- Joint Task Meta Optimization (JTMO)
- Cross Task Gradient Stabilization (CTGS)



Quantitative Results

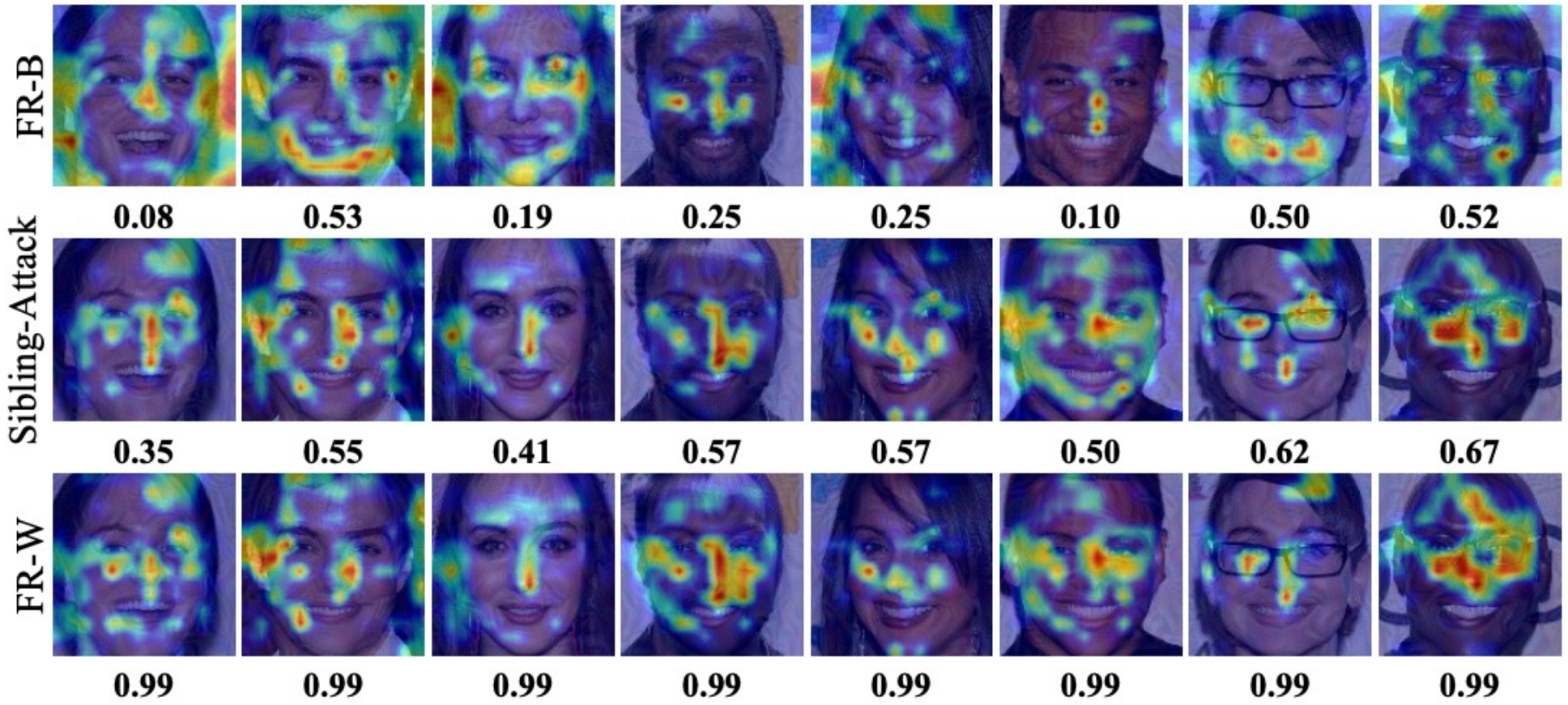
- Sibling-Attack improves the attack success rate by **12.61%** and **55.77%** on average on pre-trained face recognition models and commercial face recognition systems.

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 ↑

UC RIVERSIDE

Qualitative Results

- Gradient responses from Sibling-Attack and the target mode (FR-W) both focus more on the similar key facial regions, interprets the stronger transferability.



UC RIVERSIDE

Ablation Study & Analysis

- Attack success rate gradually increase with adding each proposed component, validating effectiveness of each components.

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**

- Sibling-Attack could generate visually-indistinguishable adversarial examples competitive to mainstream methods.

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**

*Hard P.S.: hard parameter sharing; JTMO: Joint Task Meta Optimization; CTGS: Cross Task Gradient Stabilization.

UC RIVERSIDE

Discussion & Conclusion

- Go beyond face recognition: boost transferability of attacking other tasks.
- Adversarial attack for good: improve model robustness.

- Attack success rate of Sibling-Attack significantly outperforms current SOTA single-task attacks particularly on several online commercial FR systems by a large margin.

- Related Links:

Paper

Homepage

UC RIVERSIDE