南京大学 NANJING UNIVERSITY
南京大学计算机科学与技术系
计算机软件新技术国家重点实验室(南京大学)
State Key Laboratory for Novel Software Technology at Nanjing University
RINC Robotic Intelligence & Neural Computing Group

面向人脸识别的对抗攻击及其防御研究

答 辩 人：刘小亮（2018级）
导　　师：申富饶教授 聂长海教授
研究方向：人工智能（神经网络）安全

南京大学计算机科学与技术系

2024年05月23日

南京大学
NANJING UNIVERSITY

目录
CONTENTS

一、研究背景与意义

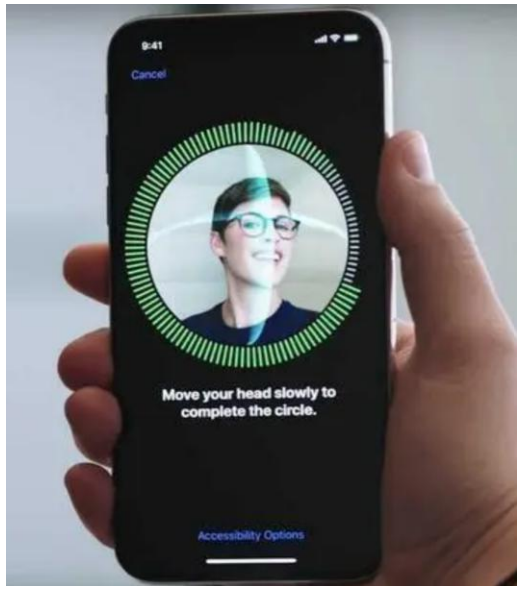
人脸识别技术的普及性与重要性

金融领域

人脸支付，远程刷脸开卡等

公共安全领域

协助抓捕嫌疑犯，抓拍交通违法，
寻找失踪人员，机场安检等

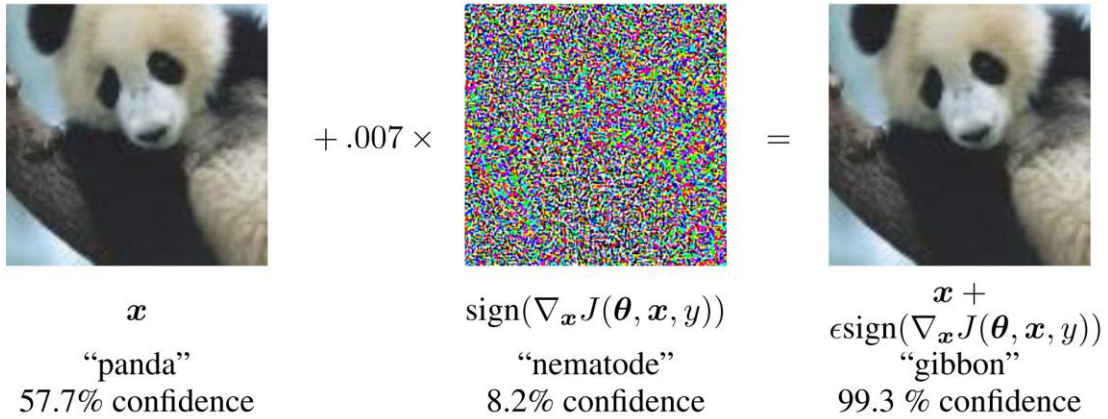


其他日常应用

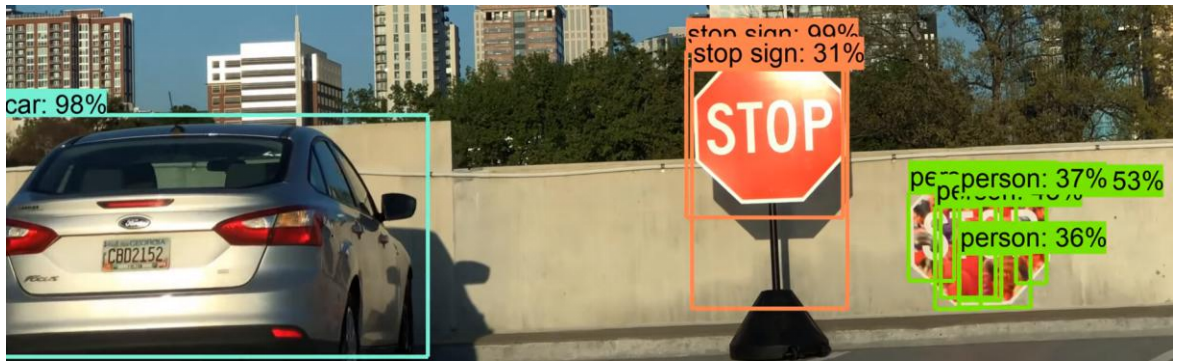
手机解锁，小区门禁，
员工考勤管理，未成年
游戏防沉迷等

对抗攻击的威胁

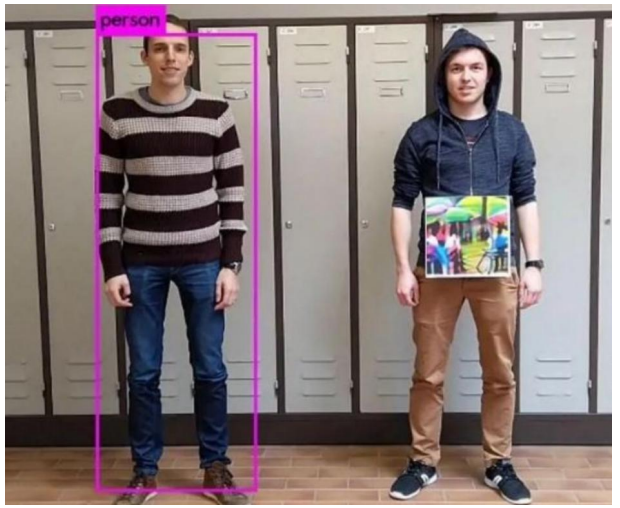
对抗扰动 对抗样本



FGSM [Goodfellow IJ, Shlens J, et al. (2014)]

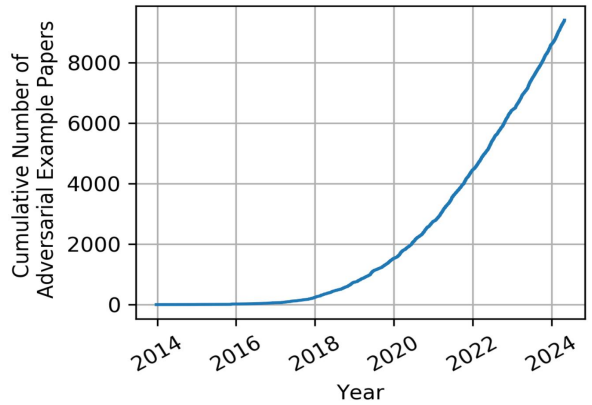


[Chen ST, Cornelius C, et al. (2019)]

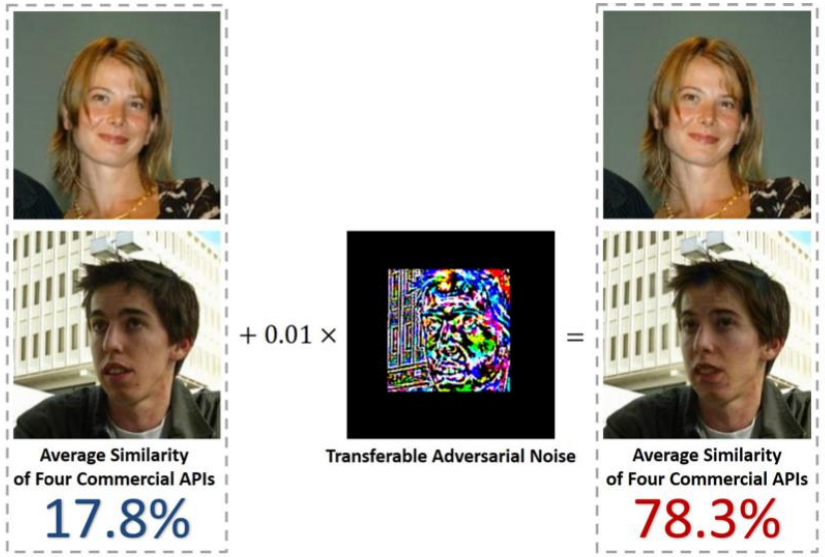


[Thys S, Van Ranst W, et al. (2019)]

对抗图块

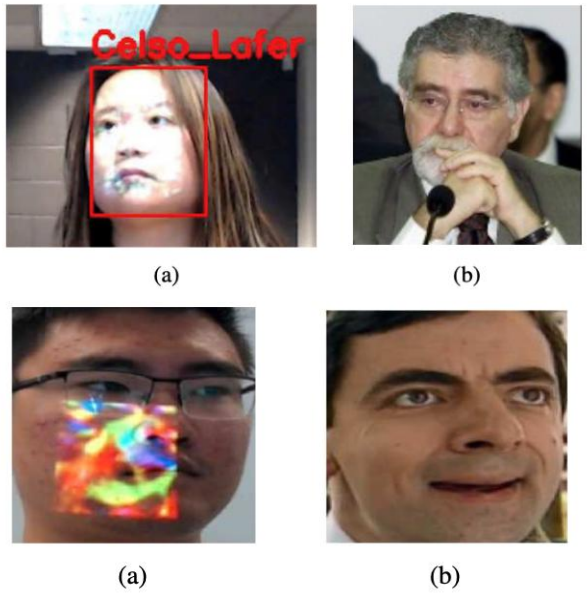


对抗攻击的威胁



DFANet-DI-M-FIM
[Zhong Y, Deng W. (2020)]

Adv-Glasses
[Sharif M, Bhagavatula S, et al. (2021)]



Adv-Light
[Nguyen DL, Arora SS, et al. (2020)]

- **增强系统安全性（主要研究意义）**

通过同时研究对抗攻击和防御机制，全面提升人脸识别系统的安全性，确保系统能够抵御各种潜在的攻击，保护关键的基础设施和敏感数据。

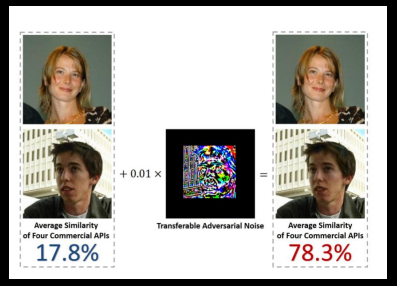
- **保障用户隐私**

人脸识别技术在个人数据保护方面扮演着重要角色。对抗攻击及其防御研究有助于保护用户的个人隐私，防止身份盗用和非法访问。

- **推动人脸识别技术的发展**

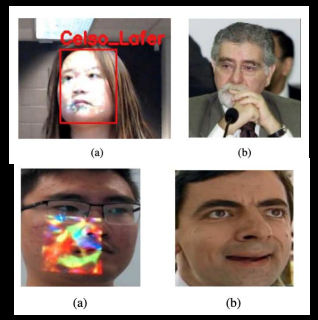
通过深入分析和解决人脸识别模型和系统中的弱点，促进人脸识别模型和系统内部逻辑的关键洞见，探索更有效的改进策略。

对抗攻击

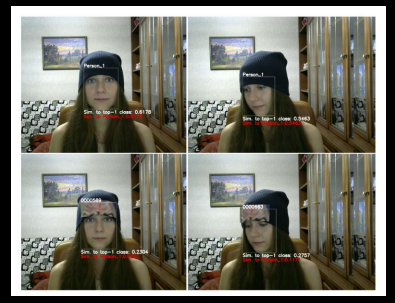


DFANet-DI-M-FIM [Zhong Y, Deng W. (2020)]

Adv-Glasses [Sharif M, Bhagavatula S, et al. (2021)]



Adv-Light [Nguyen DL, Arora SS, et al. (2020)]



Adv-Hat [Komkov, S, Petiushko A.(2021)]

对抗防御

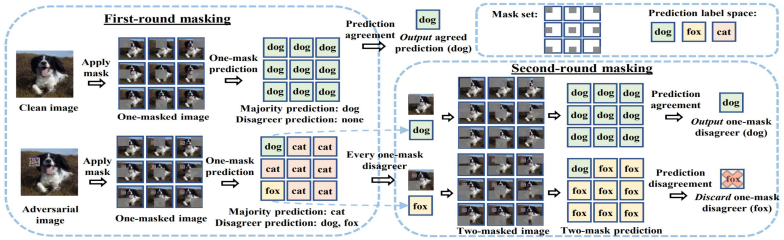
- **对抗训练**（Adversarial Training, AT）：**提升模型对抗鲁棒性。**

$$\min_{\theta} \frac{1}{|D|} \sum_{x,y \in D} \max_{\|\delta\| \leqslant \epsilon} \mathcal{L}(\mathcal{F}_{\theta}(x+\delta), y).$$

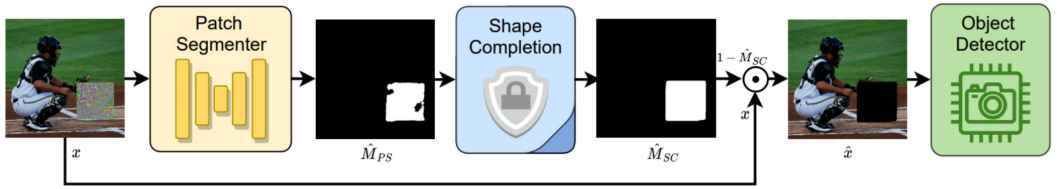
对抗风险最小化　　生成对抗样本

PGDAT [Madry A, et al.(2018)]

- **对抗图块的防御：**



判断−屏蔽图块，PatchCleanser(PC) [Chong,et al.(2022)]



检测−屏蔽图块，Segment and Complete(SAC) [Liu,et al.(2022)]

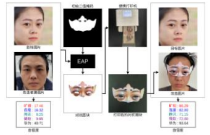
黑盒环境下的攻击低成功率问题	• 黑盒攻击者无法访问目标模型的内部信息，这一限制使得设计有效攻击变得更加困难，从而导致攻击成功率低。
三维环境下黑盒攻击困难及低成功率问题	• 在三维复杂环境下，由于空间和特征的维度增加，构造有效的黑盒攻击比在二维环境中更为困难，因此攻击成功率相对较低。
对抗训练中的对抗鲁棒性与泛化能力的平衡问题	• 当前的对抗训练方法虽然可以提升模型的对抗鲁棒性，但这通常会以降低模型泛化能力为代价。
多样化对抗图块攻击下防御的局限性问题	• 现有的防御机制在面对多样化和复杂的对抗图块攻击时，往往表现出不能自适应的防御。

面向人脸识别的对抗攻击及其防御研究

第一章 绪论
第二章 相关工作

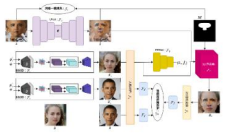
对抗攻击

科学问题一：基于黑盒对抗攻击的人脸识别模型和系统脆弱性分析



第三章 基于二维变换的黑盒对抗图块攻击方法

技术演进：维度的扩展



第四章 基于三维神经辐射场的黑盒对抗图块攻击方法

攻防研究转换

对抗防御

科学问题二：人脸识别模型的对抗鲁棒性与泛化能力之间的平衡探索

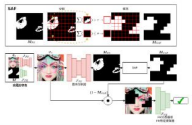
普适性与特异性的探索

科学问题三：针对多样化的对抗图块攻击自适应防御机制研究

第五章 增强模型多重鲁棒性的对抗训练防御方法

普适性与特异性的探索

第六章 自适应的对抗图块防御方法

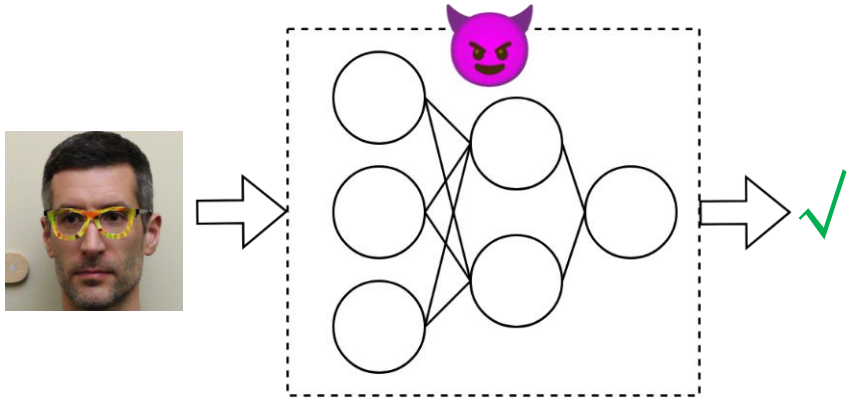


第七章 总结与展望

南京大学
NANJING UNIVERSITY

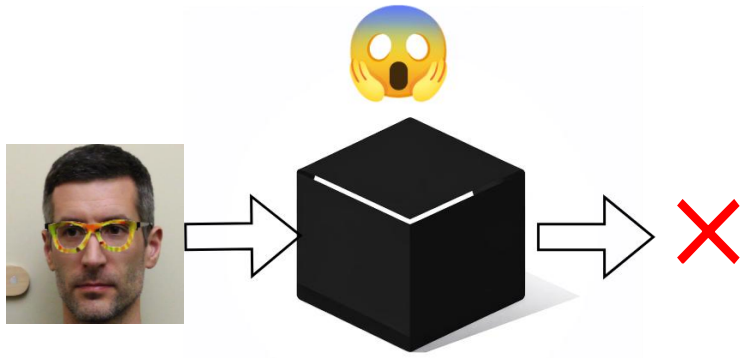
课题一、基于二维变换的黑盒对抗图块攻击方法

黑盒环境下的攻击低成功率



白盒环境

- 完全透明
- 适用范围低
- 攻击成功率高



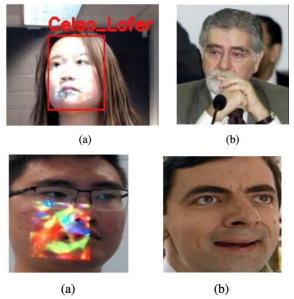
黑盒环境

- 有限知识
- 适用性更广泛
- 攻击成功率低

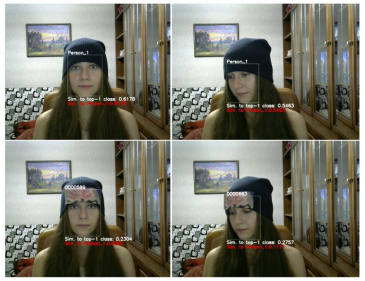
动机：提升在人脸识别下黑盒攻击的成功率

- 基于白盒的对抗图块攻击方法,在黑盒的情况下表现不佳

Adv-Glasses [Sharif M, Bhagavatula S, et al. (2021)]



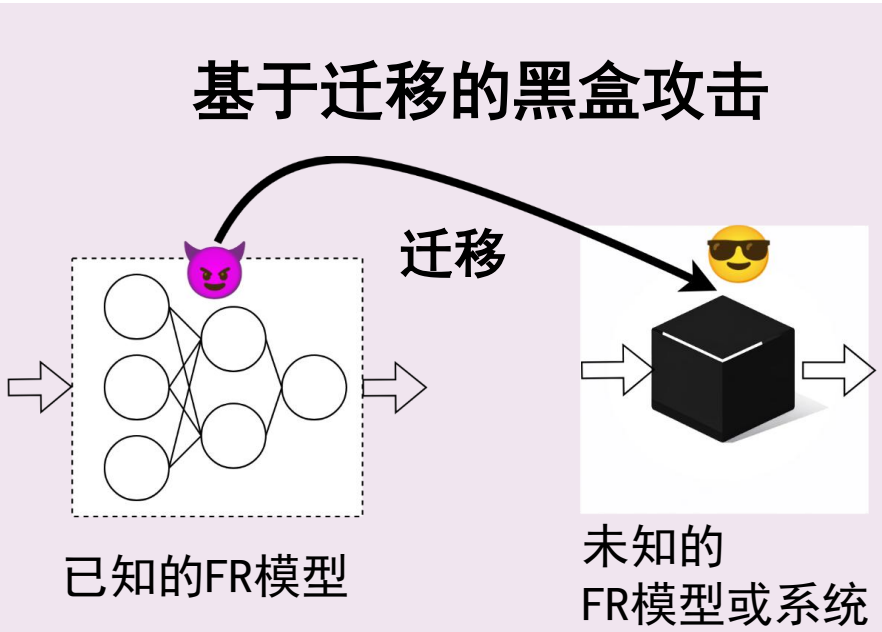
Adv-Light [Nguyen DL, Arora SS, et al. (2020)]



Adv-Hat [Komkov, S, Petiushko A.(2021)]

- 基于可迁移的对抗样本攻击

1. MIM，MIFGSM：添加动量项 稳定梯度下降的方向。
2. TIM，TIFGSM：利用图像的输入平移不变性，对梯度进行高斯滤波，降低过拟合，提升可迁移性。
3. DIM，DIFGSM：通过提升输入的多样性，降低过拟合，提升可迁移性。DIM主要是对输入图进行resize 操作。

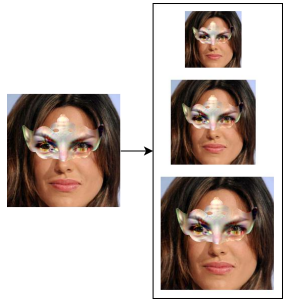


EAP主要工作：通过**提升对抗图块的可迁移性，**来增强黑盒攻击成功率

一、提升输入的多样性

1、图像金字塔（IP）策略

$$H_i = \left(H_0 + \frac{H_0}{2} \times (i-1)\right), \quad W_i = \left(W_0 + \frac{W_0}{2} \times (i-1)\right),$$



2、随机相似变换（RST）策略

1）超参数 β 获取四个维度变换参数

$$t_x = \mathcal{U}(-\beta W, \beta W),$$

$$t_y = \mathcal{U}(-\beta H, \beta H),$$

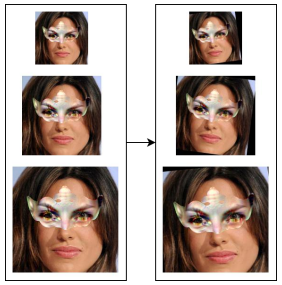
$$\theta = \mathcal{U}(-\frac{\beta\pi}{2}, \frac{\beta\pi}{2}),$$

$$s = \mathcal{U}(1-\beta, 1+\beta),$$

2）获取随机相似性变换矩阵

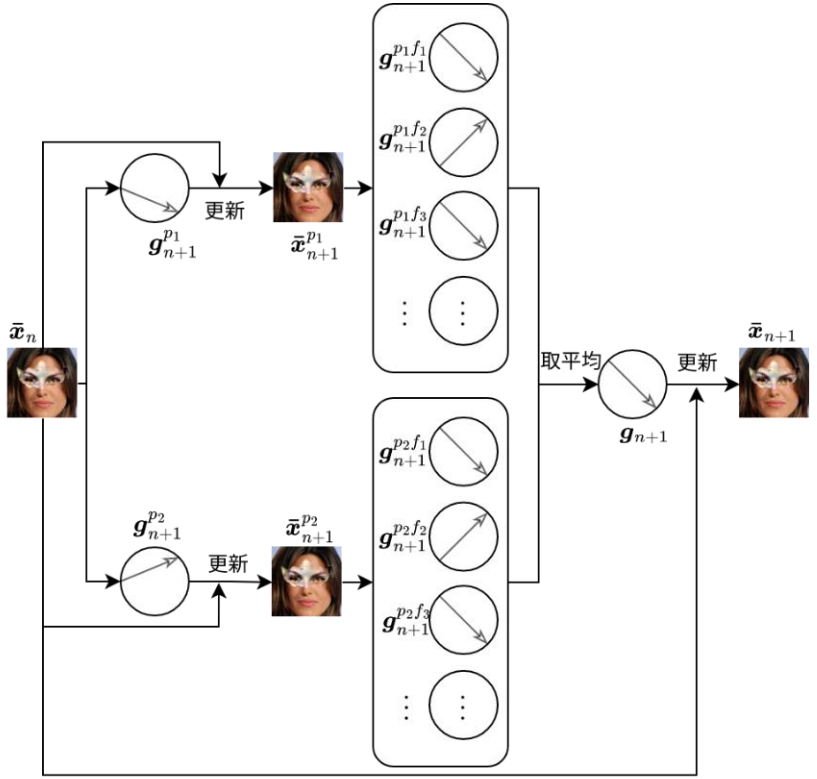
$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} s\cdot\cos(\theta) & s\cdot\sin(\theta) & t_x \\ -s\cdot\sin(\theta) & s\cdot\cos(\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix} \cdot$$



二、获取多个模型更常见的梯度特征

3、元集成攻击（MEA）策略



元集成攻击示例图（论文图3-4）

对于不同骨干网络的人脸识别模型的黑盒冒充攻击的比较结果（论文表3-1）

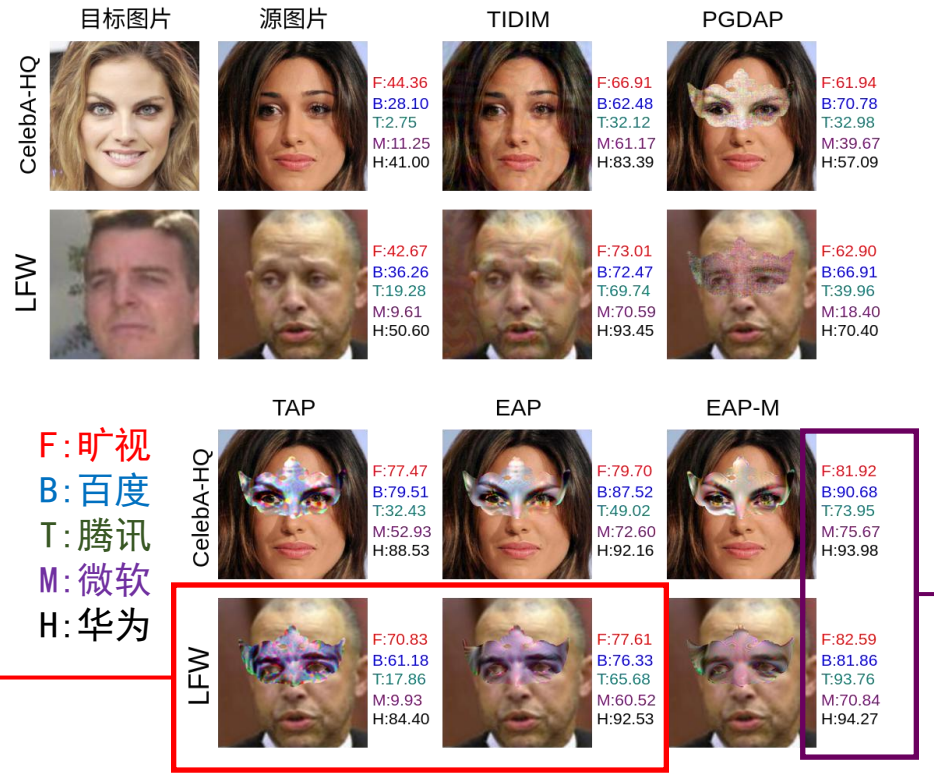
骨干网络	CelebA-HQ						LFW					
	全局方法			图块方法			全局方法			图块方法		
	PGD[11]	DIM[13]	TIDIM[14]	PGDAP[15]	TAP[16]	EAP	PGD[11]	DIM[13]	TIDIM[14]	PGDAP[15]	TAP[16]	EAP
Swin-T[83]	77.00	97.40	95.20	47.00	95.40	**99.80**	82.40	99.40	**99.80**	13.20	90.60	99.60
Effi-B0[84]	4.40	32.20	70.40	31.00	54.60	**87.20**	0.20	24.80	75.20	9.40	54.40	**85.60**
IR50[23]	2.00	37.40	79.00	24.60	56.40	**90.40**	0.00	32.00	89.20	7.20	61.40	**91.40**
IRSE152[85]	3.80	42.60	82.80	26.40	58.00	**93.40**	0.60	43.60	92.80	7.80	63.60	**93.60**
TFNAS-A[86]	4.00	45.40	81.80	38.20	54.80	**94.00**	0.00	35.40	84.80	14.40	56.60	**92.40**
ReXNetV1[87]	5.00	38.60	78.20	36.60	64.00	**94.20**	0.40	31.20	84.80	14.60	66.00	**90.60**
ResNeSt50[88]	3.20	41.00	82.40	26.80	41.40	**88.20**	0.20	33.00	**90.00**	5.60	25.40	83.40
RepVGG-A0[89]	4.60	37.00	71.20	34.00	57.60	**86.80**	0.40	36.20	86.20	18.40	72.40	**90.40**
RepVGG-B0[89]	2.20	32.40	72.60	28.40	62.60	**85.20**	0.40	39.60	89.20	13.80	76.80	**93.40**
RepVGG-B1[89]	3.60	38.40	77.40	26.60	61.00	**88.40**	0.60	40.80	90.40	12.00	75.00	**94.60**
MFNet[26]	7.00	34.80	71.00	38.00	53.20	**86.40**	0.40	23.20	75.20	17.60	54.00	**83.20**
LightCNN29[90]	3.20	25.40	60.20	40.20	60.40	**86.80**	0.00	16.80	64.80	19.80	58.20	**82.60**
HRNet[91]	3.00	40.20	82.80	25.80	54.20	**95.00**	0.60	35.40	87.20	10.60	53.20	**92.60**
GhostNet[92]	4.00	33.20	67.40	35.80	51.60	**88.60**	0.00	23.60	73.60	18.00	48.40	**85.20**
Att56[93]	4.00	52.20	90.40	37.60	74.80	**98.20**	1.00	52.60	96.80	15.60	75.60	**97.40**
Att92[93]	2.80	45.60	86.20	25.60	67.40	**97.00**	0.40	47.60	94.20	10.00	70.60	**96.20**
平均值	8.36	42.11	78.06	32.66	60.46	**91.22**	5.48	38.45	85.89	13.00	62.64	**90.76**

高30.76%　　高28.12%

1. 在CelebA-HQ上，EAP在所有16个骨干网络上都获得了最佳攻击性能。
2. 平均攻击成功率在CelebA-HQ上比TAP高30.76%，在LFW上高28.12%。

对SOTA商业人脸识别系统进行黑盒冒充攻击的比较结果（论文表3-1）

		旷视[102]	百度[103]	腾讯[104]	微软[105]	华为[106]	平均值
数据集：CelebA-HQ							
全局方法	PGD[11]	45.18	31.06	15.46	12.09	50.73	30.91
	DIM[13]	57.37	47.69	28.45	28.98	72.28	46.95
	TIDIM[14]	66.32	58.08	40.05	42.09	83.63	58.03
图块方法	PGDAP[15]	61.77	53.07	24.51	20.77	64.63	44.95
	TAP[16]	65.04	54.97	22.17	20.35	66.49	45.80
	EAP	70.25	65.77	34.58	37.84	78.96	57.48
	EAP-H	75.76	73.75	50.47	54.47	90.11	68.91
	EAP-M	**75.94**	**73.89**	**50.69**	**54.67**	**90.23**	**69.08**
数据集：LFW							
全局方法	PGD[11]	38.03	21.87	15.25	10.96	48.33	26.89
	DIM[13]	52.56	43.30	28.38	31.72	72.29	45.65
	TIDIM[14]	64.78	61.61	43.58	53.79	84.95	61.74
图块方法	PGDAP[15]	54.64	41.15	20.70	14.83	55.52	37.37
	TAP[16]	60.18	50.86	19.56	16.50	61.69	41.76
	EAP	66.12	62.90	32.68	36.21	76.87	54.96
	EAP-H	71.88	71.97	50.07	54.28	**89.54**	67.55
	EAP-M	**71.92**	**72.12**	**51.14**	**54.46**	89.49	**67.83**



F：旷视
B：百度
T：腾讯
M：微软
H：华为

1. 与EAP相比，集成攻击可以带来攻击成功率较大的提升。
2. 相比于EAP-H，EAP-M提升更高。
3. 攻击“华为”时平均置信度都在90%左右。

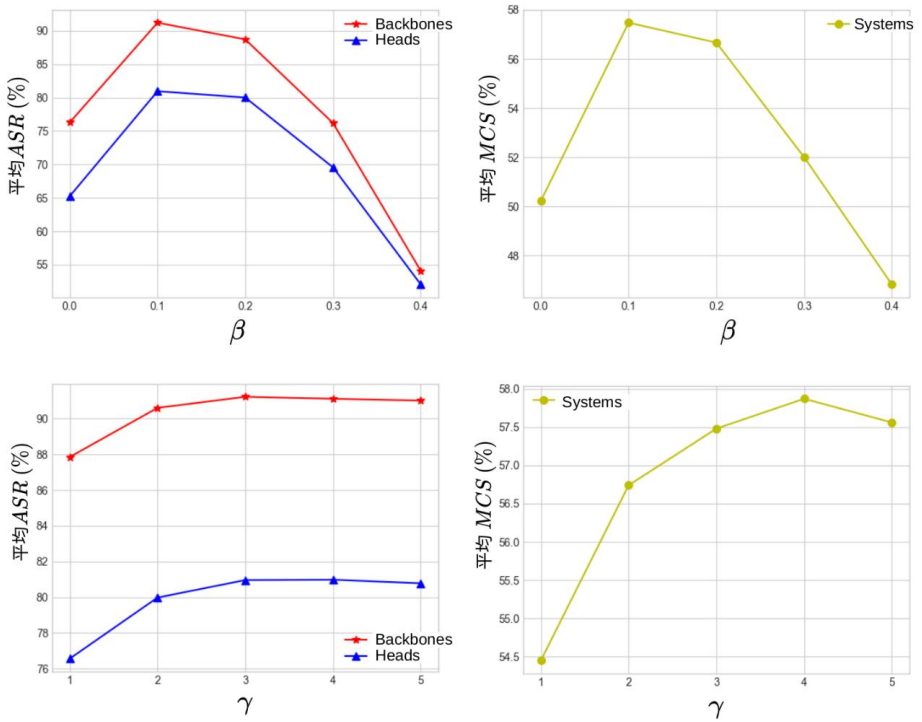
4. TAP vs EAP：腾讯：17.86 vs 65.68(↑47.82)，微软：9.93 vs 60.52(↑50.59)。

5. 在五个商业系统中，EAP-M的平均置信度都高于70%。

消融实验结果。随机相似变换（RST），图像金字塔（IP），元集成攻击（MEA）（论文表3-5）

RST	IP	MEA	骨干网络	损失函数	系统
✘	✘	✘	60.46	52.67	45.80
✔	✘	✘	87.85	76.56	54.45
✘	✔	✘	76.28	65.24	50.23
✔	✔	✘	**91.22**	**80.96**	**57.48**
✘	✘	✔	-	-	65.62
✔	✘	✔	-	-	66.63
✘	✔	✔	-	-	67.52
✔	✔	✔	-	-	**69.08**

1. 三个策略都带来性能的提升。
2. 相比IP策略，RST策略提升更大。
3. 三种策略联合使用时，攻击效果显著优于单独应用各策略的效果。



超参数敏感性实验结果，β控制相似变换尺度的，γ控制的图像金字塔级数（论文图3-8）

4. 合适β值显著增强对抗图块的可迁移性，并提升攻击效果。较高的β值可能导致对抗图块过度调整，从而降低攻击性能。
5. γ增大到一定程度，性能基本不再增加。

EAP的核心贡献

- **创新攻击方法**：提出了基于二维变换的可迁移对抗图块攻击方法。
- **策略综合应用**：提出随机相似变换、图像金字塔处理及元集成攻击策略，显著提升了攻击的可迁移性。
- **广泛验证**：在多个先进人脸识别模型及商业系统上的实验验证，展现其高效和实用性。

进一步工作方向

- **扩展到三维环境**：当前研究集中于二维攻击，进一步需要探索更为复杂的三维场景中的攻击方法。

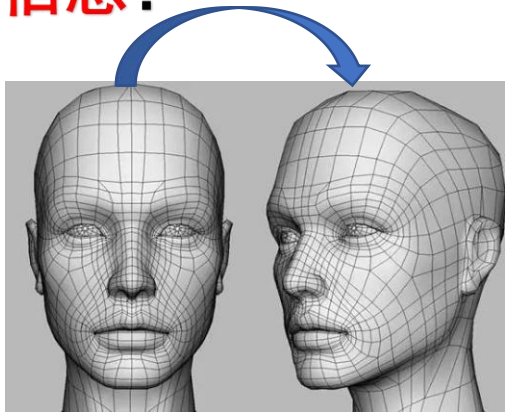
南京大学
NANJING UNIVERSITY

课题二、基于三维神经辐射场的黑盒对抗图块攻击方法

三维环境攻击的困难性

单张图获取三维信息？

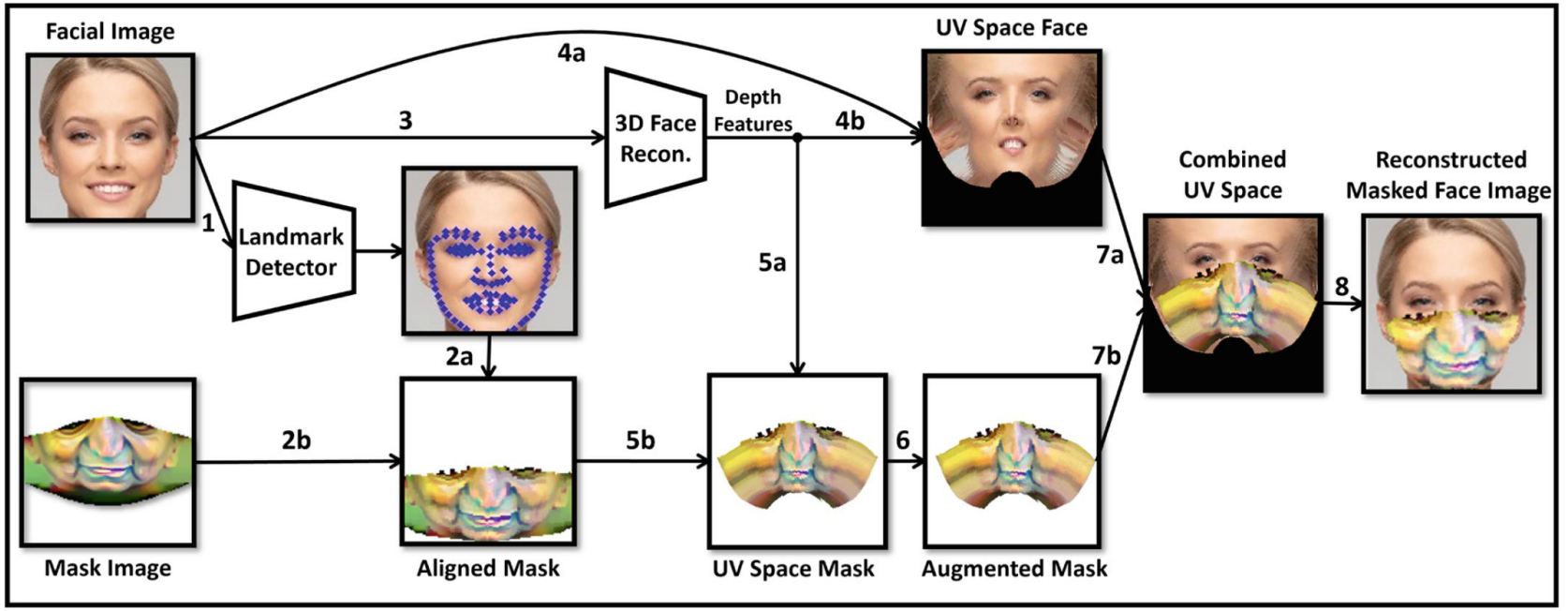
获取到的目标人脸图
（**已知的**）



人脸三维信息

人脸识别系统
记录的目标人脸图
（**未知的**）

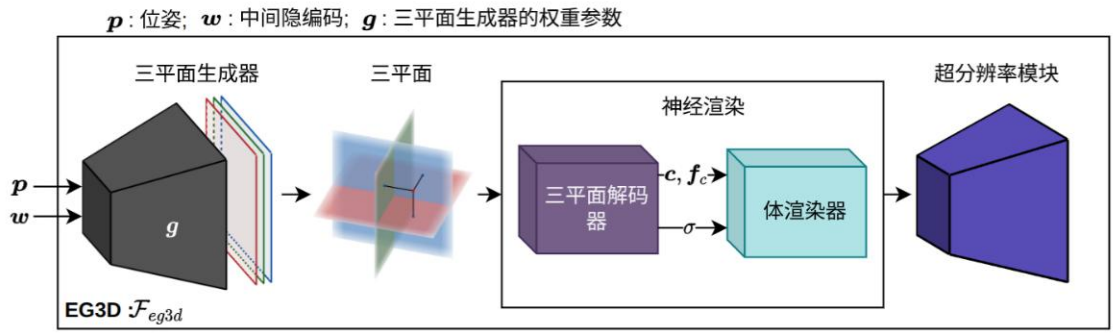
动机：利用获取三维信息提升黑盒攻击的成功率



Adversarial mask [Zolfi, A., Avidan, S., et al.(2022)]

- 扩展的到三维攻击，提升多视角下攻击的成功率。
- 主要针对白盒的攻击，黑盒攻击下成功率低。
- 人脸RGB信息还只是原来二维图像的。

基于NeRF的3D-GAN逆映射



EG3D框架图（论文图4-2）

步骤一：优化中间隐编码

$$\boldsymbol{w}_x, \boldsymbol{n} = \underset{\boldsymbol{w},\boldsymbol{n}}{\operatorname{argmin}} \mathcal{J}_{LPIPS}(\boldsymbol{x}, \mathcal{F}_{eg3d}(\boldsymbol{p}, \boldsymbol{w}, \boldsymbol{g})) + \lambda_n \mathcal{J}_n(\boldsymbol{n}),$$

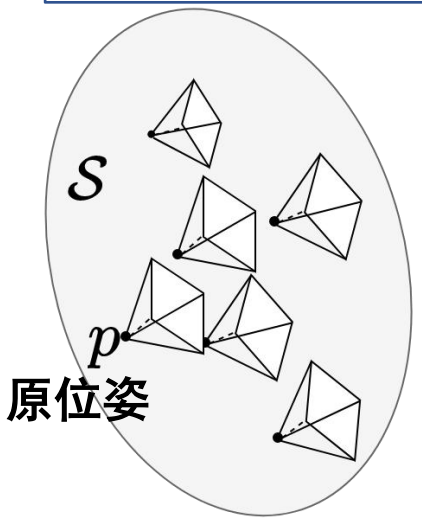
步骤二：微调生成器的参数

$$\min_{\boldsymbol{g}} \mathcal{J}_{LPIPS}(\boldsymbol{x}, \mathcal{F}_{eg3d}(\boldsymbol{p}, \boldsymbol{w}_x, \boldsymbol{g})) + \lambda_{L2} \mathcal{J}_{L2}(\boldsymbol{x}, \mathcal{F}_{eg3d}(\boldsymbol{p}, \boldsymbol{w}_x, \boldsymbol{g})),$$

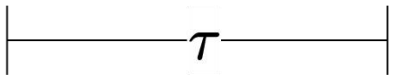
3D-GAN逆映射一个示例

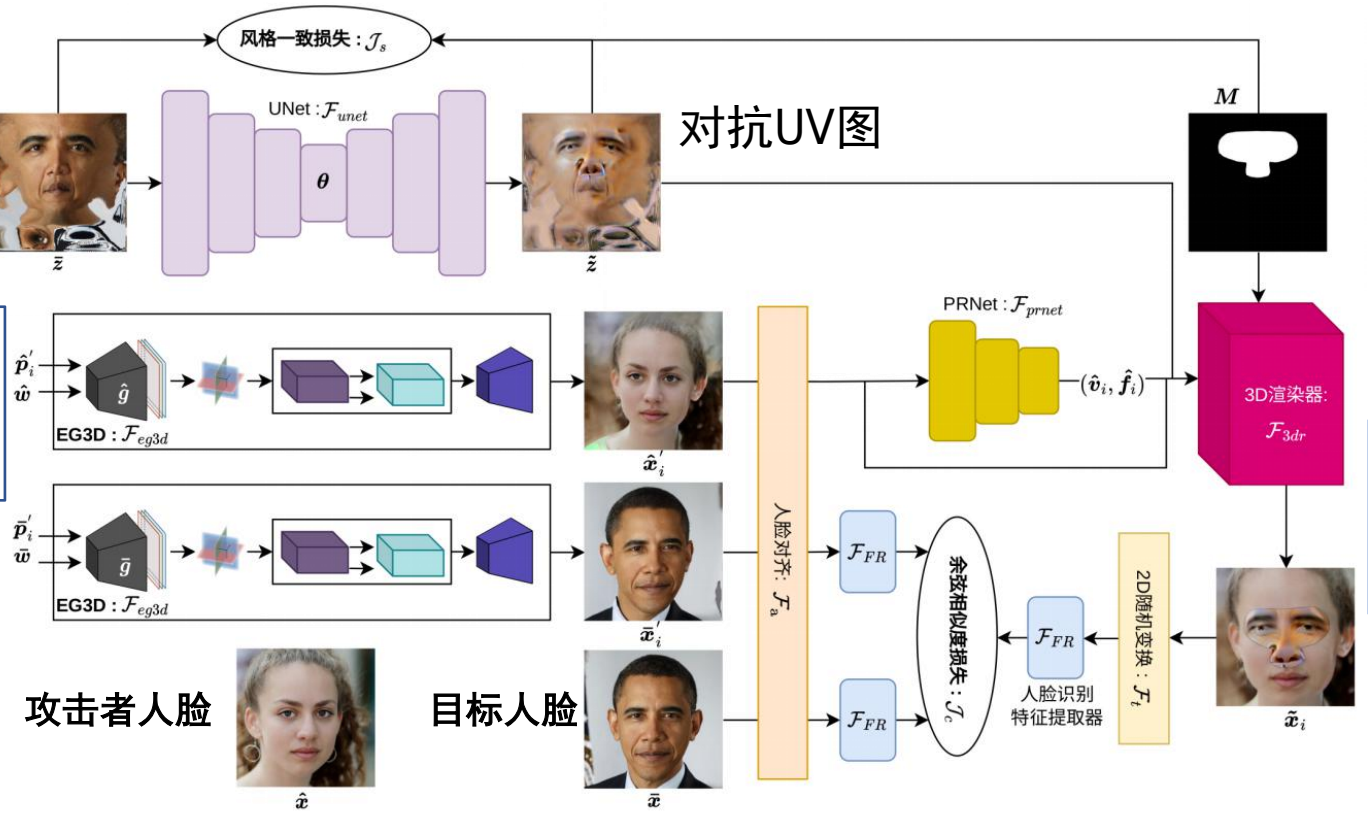
UV图：可编辑，用于生成对抗UV图。

NeRF：不容易编辑，用于生成高清的不同视角的图。



原位姿





通过对抗UV图获取相应的对抗图块。

NeRFTAP的总体框架。（论文图4-3）

优化目标：

$$\min_{\theta} \mathbb{E}_{\hat{p}_i' \sim \hat{S}, \bar{p}_i' \sim \bar{S}} \left[\mathcal{J}_c\left(\mathcal{F}_{FR}(\mathcal{F}_t(\tilde{x}_i)), \mathcal{F}_{FR}(\mathcal{F}_a(\bar{x}_i')), \mathcal{F}_{FR}(\mathcal{F}_a(\bar{x}))\right) + \lambda_s \mathcal{J}_s(\bar{z}, \tilde{z}, \boldsymbol{M}) \right],$$

$$\text{s.t.} \quad \hat{S} = \{\hat{p} + \tau \cdot (2\beta - 1) | \beta \sim \text{Beta}(\alpha, \alpha)\},$$

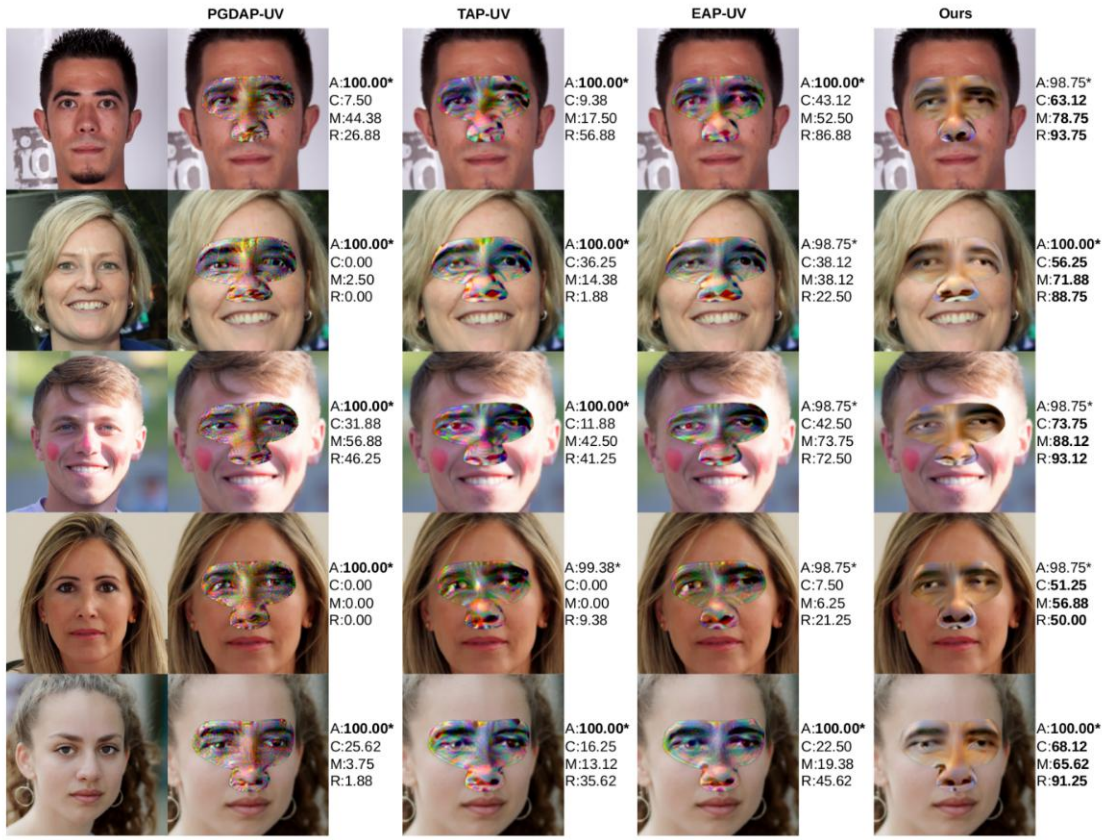
$$\bar{S} = \{\bar{p} + \tau \cdot (2\beta - 1) | \beta \sim \text{Beta}(\alpha, \alpha)\}.$$

不同场景下目标人脸图像的示例（论文图4-4）

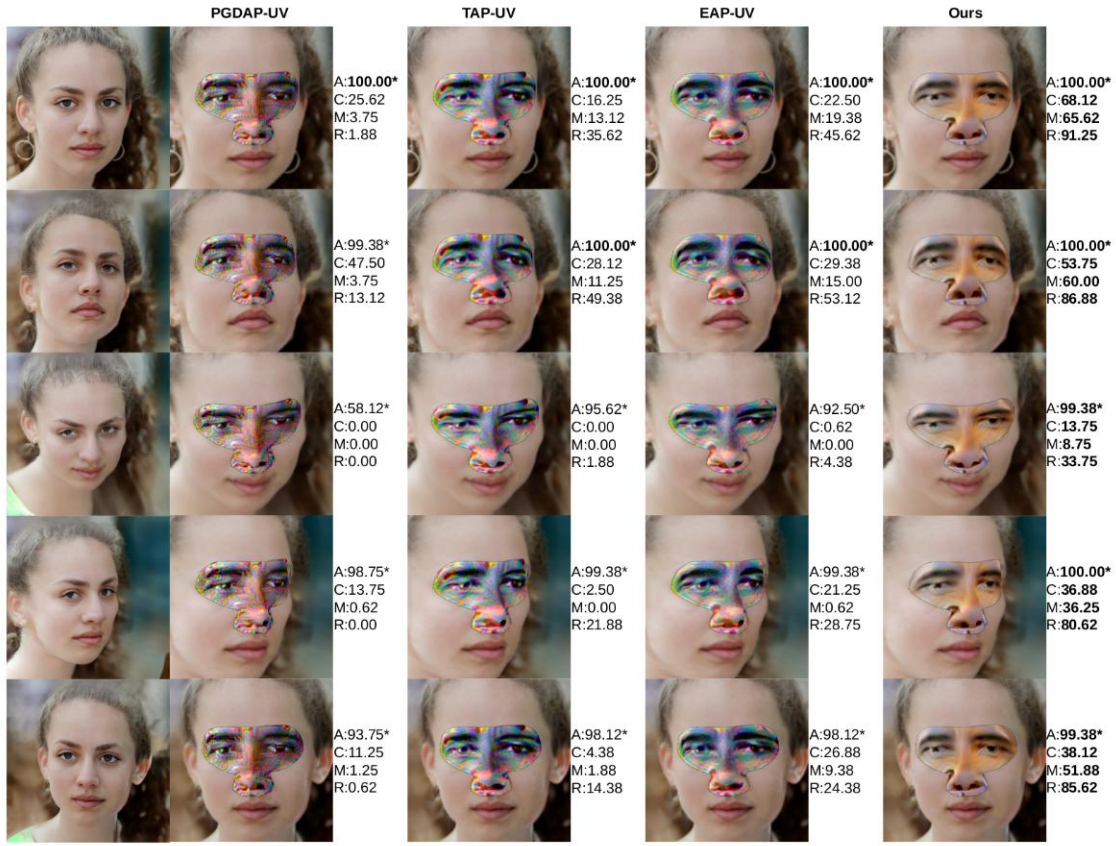
1. NeRFTAP相对于EAP-UV, ASR提升高达47.38%。
2. 同样在“腾讯”系统，MSC提升高达20.83%。

NeRFTAP与其他攻击方法的比较结果（论文表4-1）

训练时 FR 模型	方法		测试 FR 模型 Arc.[25]	Cos.[119]	Mob.[26]	Res.[23]	测试 FR 系统 旷视[102]	腾讯[104]
ArcFace[25]	全局	FGSM[10]	87.06*	2.75	27.18	23.06	52.93	34.77
		MIM[12]	**100.00***	4.00	35.62	27.50	53.29	28.92
		DIM[13]	**100.00***	19.56	64.81	50.93	63.54	38.59
		TIP-IM[109]	**100.00***	23.50	62.43	58.00	64.48	40.48
	眼部&鼻子	PGDAP-UV[15,57]	99.88*	10.88	21.75	10.19	64.05	32.90
		TAP-UV[16,57]	99.81*	10.69	23.38	23.88	66.03	31.41
		EAP-UV[17,57]	99.00*	23.62	45.81	48.75	66.96	35.14
		Ours	99.00*	**60.87**	**70.25**	**77.43**	**69.80**	**55.97**
CosFace[119]	全局	FGSM[10]	6.18	38.31*	11.75	12.75	46.39	29.40
		MIM[12]	12.75	**99.06***	13.31	11.81	47.82	26.98
		DIM[13]	38.18	98.00*	50.12	30.68	57.47	34.89
		TIP-IM[109]	44.43	98.50*	51.12	37.93	60.00	37.38
	眼部&鼻子	PGDAP-UV[15,57]	5.56	98.75*	17.56	13.81	63.70	35.96
		TAP-UV[16,57]	13.38	98.38*	35.88	31.88	63.83	37.49
		EAP-UV[17,57]	43.12	93.00*	64.19	64.75	66.15	47.19
		Ours	**90.50**	94.06*	**86.50**	**77.94**	**68.36**	**54.08**
MobileFace[26]	全局	FGSM[10]	23.93	2.62	92.81*	38.37	51.85	33.38
		MIM[12]	33.00	5.00	**99.87***	47.37	49.90	28.12
		DIM[13]	56.18	35.31	99.43*	76.25	56.37	37.67
		TIP-IM[109]	50.93	34.75	99.37*	76.18	59.31	38.18
	眼部&鼻子	PGDAP-UV[15,57]	32.19	29.62	99.38*	44.75	66.97	41.22
		TAP-UV[16,57]	69.62	43.31	98.50*	72.94	70.44	45.22
		EAP-UV[17,57]	79.12	57.38	96.69*	82.12	72.50	48.97
		Ours	**89.62**	**65.06**	98.38*	**91.31**	**75.96**	**58.09**



可视化比较结果。（论文图4-7）



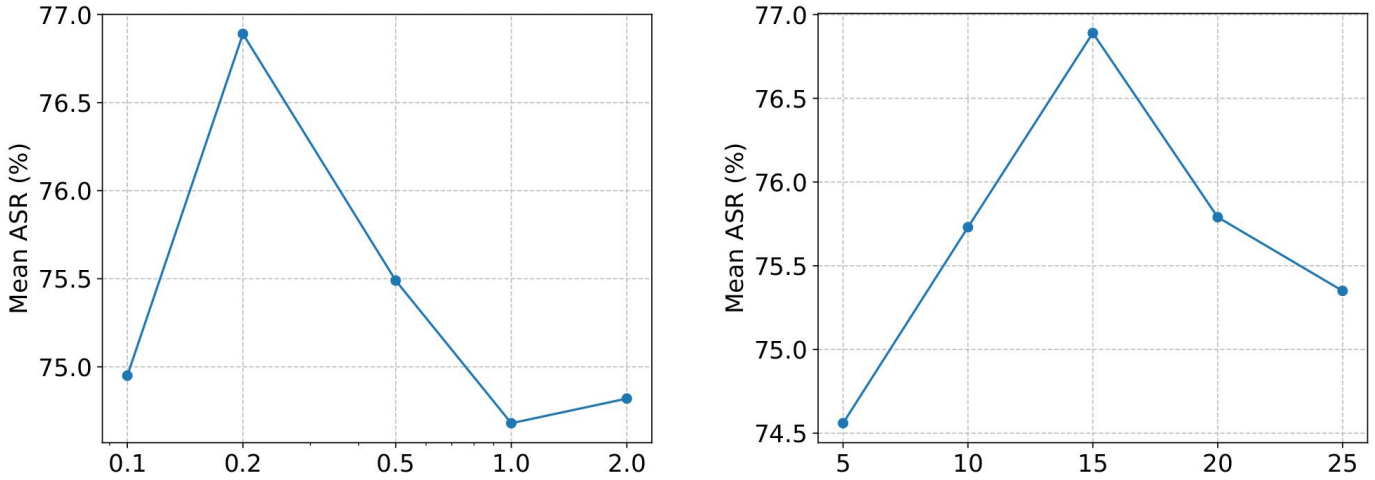
不同视角攻击者图像的比较结果。（论文图4-8）

1、在黑盒攻击设置下，NeRFTAP的性能明显超过了其他方法。
2、在不同视角源图像，NeRFTAP的性能同样表现更高的ASR。

消融实验结果。＊表示白盒攻击（论文表4-3）

	Arc.[25]	Cos.[119]	Mob.[26]	Res.[23]
全部模块	99.00*	**60.87**	**70.25**	**77.43**
w/o 风格一致损失	99.43*	43.25	49.00	52.37
w/o 2D 随机变换	**99.68***	39.37	58.00	65.62
w/o EG3D 模块	99.31*	45.62	54.12	65.75

1、“2D随机相似变换”同样在NeRFTAP中起的重要作用。
2、基于NeRF的3D-GAN生成不用视角的人脸图片可以有较的提升攻击的成功率和可迁移性。
3、“全部模块”表现最好，各模块之间协同作用，无明显相互排斥现象。



(a) 超参数 α　　(b) 缩放因子 τ (°)

超参数分析结果（论文图4-9）

4、当超参数 α 设置为 0.2，且缩放因子 τ 调整至 15度时，NeRFTAP方法展现出最佳的攻击性能。
5、缩放因子τ太小会导致视角范围有限，而太大会超出边界，产生更多噪声，相应的性能便会下降。

NeRFTAP的核心贡献

- **3D环境的攻击方法：**提出了一个基于神经辐射场对抗图块攻击方法。
- **提高攻击适用性和效果：**通过基于NeRF的3D-GAN生成多视角人脸图像，显著提升对抗图块的适用性和攻击效果。
- **新测试数据集和广泛验证：**引入新的测试数据集，并对多种人脸识别模型和系统进行细致的实验和评估，验证了NeRFTAP的明显攻击优势。

进一步工作方向

- **防御策略的开发：**提升人脸模型对抗鲁棒性，与防御对抗图块攻击的研究。

南京大学
NANJING UNIVERSITY

课题三、增强模型多重鲁棒性的对抗训练防御方法

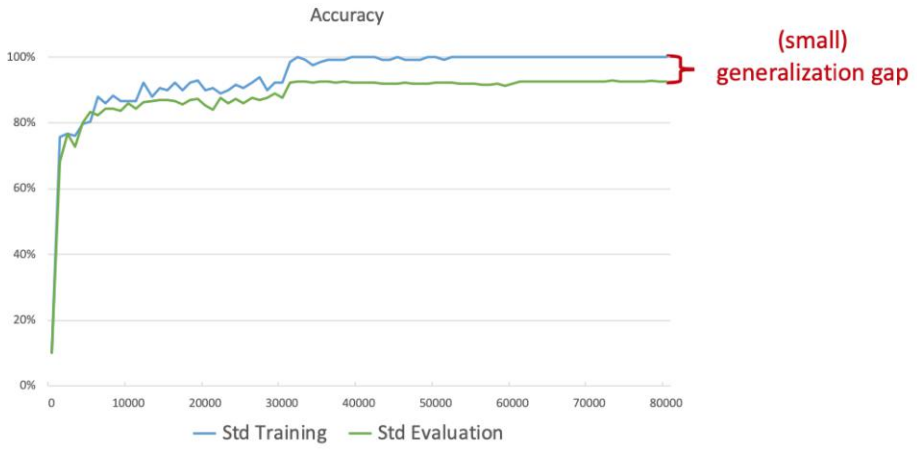
模型的对抗鲁棒性与泛化能力的平衡问题

对抗训练：$\min_{\theta} \frac{1}{|D|} \sum_{x,y \in D} \max_{\|\delta\| \leqslant \epsilon} \mathcal{L}(\mathcal{F}_{\theta}(x+\delta), y)$

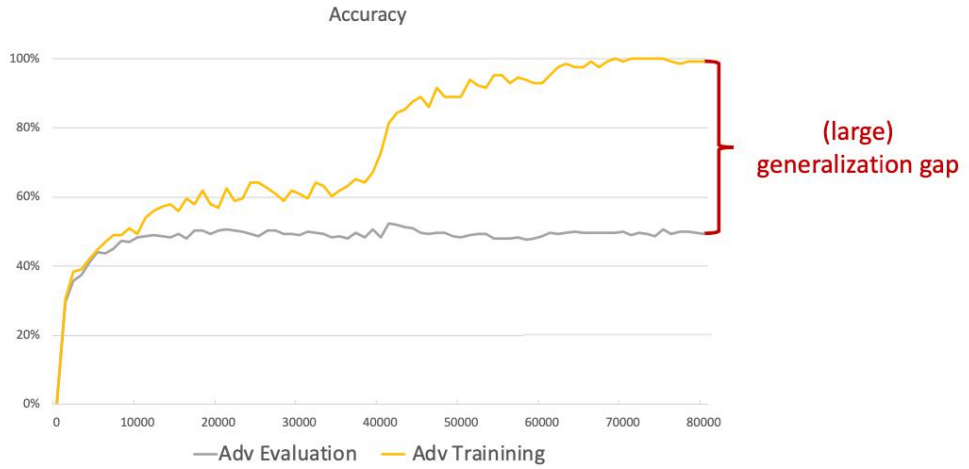
生成对抗样本

最小化 ⟷ 最大化

此消彼长

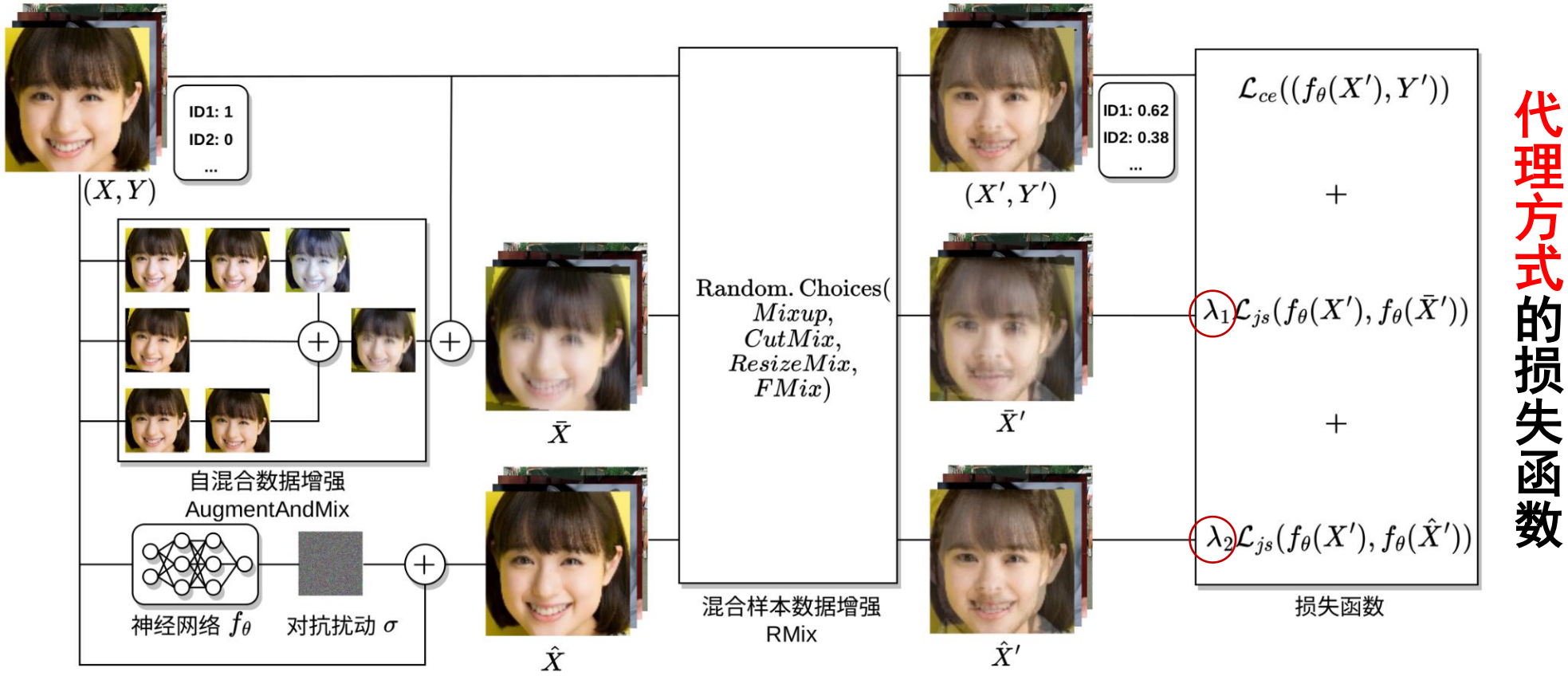


标准训练



PGDAT对抗训练

动机：平衡模型的对抗鲁棒性和泛化能力



代理方式的损失函数

AugRmixAT总体框架图（论文图5-1）

- **数据增强：**结合数据增强方法和对抗训练，增加模型对不同类型变化的适应能力。
- **调整对抗样本的强度：**通过代理方式来训练对抗样本，利用超参数来调整对抗样本的占比强度

在人脸数据集上，白盒攻击比较结果（论文表5-1）

比较方法	Clean	白盒攻击 FGSM	PGD10	PGD20	CW20
Standard	88.47	20.48	0.00	0.00	0.00
Mixup[62]	88.61	1.49	0.00	0.00	0.00
CutMix[127]	90.98	7.48	0.00	0.00	0.00
AugMix[133]	91.98	24.96	0.00	0.00	0.00
PGD-AT[11]	72.09	63.44	26.74	24.63	24.42
TRADES-1[31]	74.71	66.89	33.45	31.73	29.83
TRADES-6[31]	67.60	61.35	37.75	36.94	32.74
IAT[61]	82.12	74.45	31.80	28.79	22.83
Ours-8-1	**93.38**	**79.99**	10.70	7.21	1.87
Ours-8-8	85.05	78.67	42.55	40.56	29.08
Ours-8-32	77.44	71.31	**45.64**	**44.72**	**34.70**

迁移黑盒攻击比较结果（论文表5-2）

防御模型	攻击模型 Standard	PGD-AT	TRADES-1	IAT
Standard	-	80.64	79.00	69.70
Mixup[62]	58.81	81.16	79.39	67.24
CutMix[127]	71.18	86.04	85.34	78.72
AugMix[133]	83.33	86.05	84.12	77.08
PGD-AT[11]	69.83	-	57.45	59.17
TRADES-1[31]	72.77	62.23	-	61.90
TRADES-6[31]	65.60	56.53	55.26	57.32
IAT[61]	79.64	70.89	68.61	-
Ours-8-1	**89.43**	**87.99**	**86.67**	**80.81**
Ours-8-8	83.04	75.94	73.60	70.73
Ours-8-32	75.17	67.17	64.66	63.58

1. 其他对抗训练方法都会牺牲 Clean 准确率，而 “Ours-8-1” 相比于 “Standard”，Clean 准确率不仅没有下降还提升了 4.91%，达到了最好，且黑盒鲁棒性达到了最好。
2. “Ours-8-32”整体白盒对抗鲁棒性达到最好。

常见腐化鲁棒性的评估结果（论文表5-3）

腐化/方法		Standard	Mixup[62]	CutMix[127]	AugMix[133]	PGD-AT[11]	TRADES-1[31]	TRADES-6[31]	IAT[61]	Ours-8-1	Ours-8-8	Ours-8-32
扰动	高斯	61.78	66.71	45.06	56.01	59.95	62.96	50.96	65.62	**74.96**	73.46	66.28
	散粒	55.69	62.04	41.74	56.21	53.70	55.91	44.51	59.69	**73.85**	71.03	63.54
	脉冲	60.34	67.76	47.16	56.38	53.43	58.86	44.29	61.42	**75.09**	72.49	64.94
	斑点	61.49	67.07	53.54	70.32	55.96	60.18	48.10	63.77	**81.61**	76.98	69.50
模糊	散焦	48.85	59.99	46.30	77.51	47.57	49.12	45.28	58.27	**80.43**	68.68	59.21
	玻璃	64.69	68.53	65.38	81.27	53.74	56.81	51.02	65.80	**85.17**	74.76	64.52
	运动	59.57	62.36	59.57	**82.19**	44.18	47.17	41.30	54.76	81.72	69.15	58.28
	缩放	77.44	80.30	75.75	**88.26**	59.55	62.95	55.54	71.99	87.53	75.79	65.67
	高斯	56.18	67.05	53.06	79.60	50.64	52.09	47.62	60.62	**83.10**	70.74	61.32
天气	雪	63.70	61.70	63.57	**73.09**	43.45	45.52	37.32	63.00	69.75	57.59	46.71
	霜	53.89	65.26	54.22	**66.61**	30.45	30.74	26.90	47.85	66.37	42.92	28.92
	雾	50.74	**70.77**	63.35	70.47	6.57	8.30	6.41	9.43	64.16	24.19	14.78
	雨	63.34	70.76	77.16	80.10	53.65	54.93	49.54	61.22	**81.10**	72.32	64.77
	强光	72.18	76.24	77.21	**86.77**	34.16	39.74	29.92	58.08	83.64	64.21	48.38
数字	对比度	34.72	59.02	43.50	**61.70**	7.34	7.41	6.63	8.12	53.78	18.16	11.49
	弹性变换	70.44	69.28	78.69	**89.49**	53.88	58.40	52.01	64.81	88.88	76.48	65.81
	像素化	86.98	87.32	83.67	90.69	69.59	72.08	64.97	80.18	**91.56**	83.46	75.49
	饱和度	77.81	81.12	84.00	**84.51**	70.94	70.46	66.07	77.77	83.20	71.64	66.03
	JPEG	86.26	84.85	87.56	91.73	70.93	73.67	66.37	81.21	**92.07**	84.39	76.49
mCA		63.48	69.90	63.18	75.94	48.41	50.91	43.93	58.61	**78.84**	65.71	56.43

2、对抗训练，对“扰动”鲁棒。

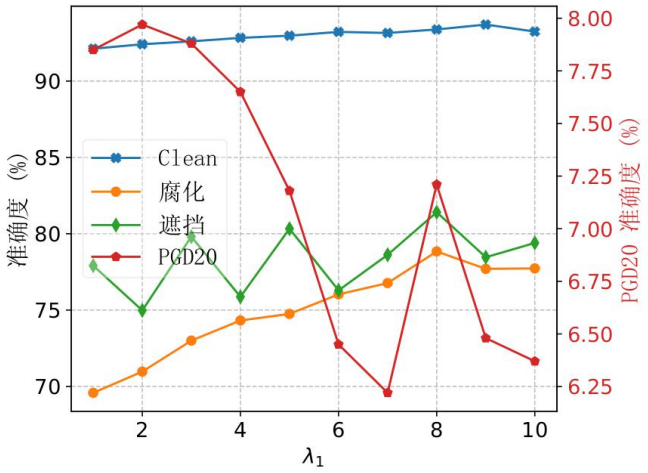
3、对抗训练，对“雾”与“对比度”相对敏感。

1、AugMix 75.94% vs 78.84% Ours-8-1(提升2.90%)

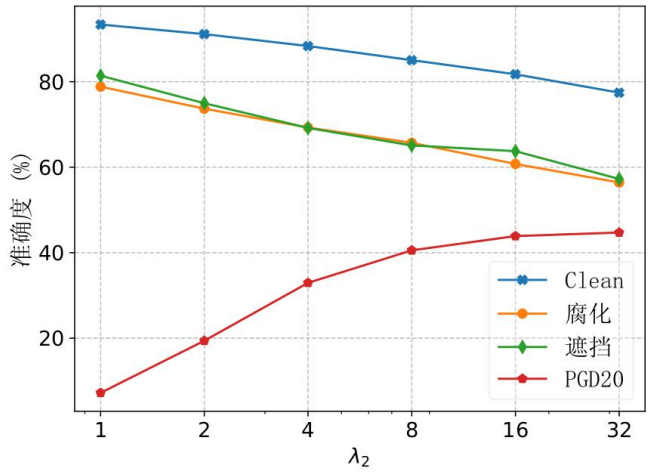
遮挡鲁棒性评估结果（论文表5-4）

比较方法	无目标	有目标	平均值
Standard	59.18	70.26	64.72
Mixup[62]	60.26	74.08	67.17
CutMix[127]	**76.95**	84.99	80.97
AugMix[133]	53.77	81.47	67.62
PGD-AT[11]	33.94	52.61	43.27
TRADES-1[31]	31.02	57.98	44.50
TRADES-6[31]	21.67	51.18	36.43
IAT[61]	28.69	62.50	45.60
Ours-8-1	75.24	**87.58**	**81.41**
Ours-8-8	58.95	71.25	65.10
Ours-8-32	42.78	71.72	57.25

1. PGD-AT，TRADES与IAT对抗训练方法可能无法有效处理非对抗性质的遮挡问题。
2. 而AugRmixAT相对来说，可以很好处理遮挡问题。



(a) 超参数 λ_1，其中 $\lambda_2 = 1$



(b) 超参数 λ_2，其中 $\lambda_1 = 8$

超参数λ_1和λ_2敏感性分析折线图（论文图5-4）

4. 更高的 λ_1 值有助于提升模型在非对抗性扰动下的腐化鲁棒性。
5. 更高的 λ_2 值增强了模型的对抗鲁棒性，但同时可能牺牲了模型正常的泛化性能。
6. 适当的参数设置可以在各方面之间实现有效的平衡。

AugRmixAT的核心贡献

- **提出了AugRmixAT对抗训练框架，**解决了对抗鲁棒性和泛化能力的平衡问题。
- **提升多重鲁棒性：**AugRmixAT增强了模型在对抗性攻击、遮挡以及常见图像腐化的处理能力，提高了在复杂环境中的可靠性。
- **泛化性能的保持：**利用软交叉熵和JS散度一致性损失，AugRmixAT在增强模型多重鲁棒性的同时保证了原有的泛化性能。

进一步工作方向

- **针对对抗图块的防御研究：**对抗训练方法可以提升人脸模型整体对抗鲁棒性，但对于对抗图块的防御力不足。进一步将针对复杂对抗图块的防御策略进行研究。

防御方法	Clean	掩码				
		眼镜	贴纸	口罩	R-mask	F-mask
Standard	89.43	0.00	0.04	0.00	14.12	5.51
Cutout[123]	92.97	0.00	0.05	0.00	13.85	5.74
AugRmixAT-8-1	**93.38**	1.90	15.32	0.34	18.23	10.45
AugRmixAT-8-32	77.44	**8.94**	**25.35**	**5.64**	**37.93**	**18.45**

AugRmixAT防御对抗图块实验结果

南京大学
NANJING UNIVERSITY

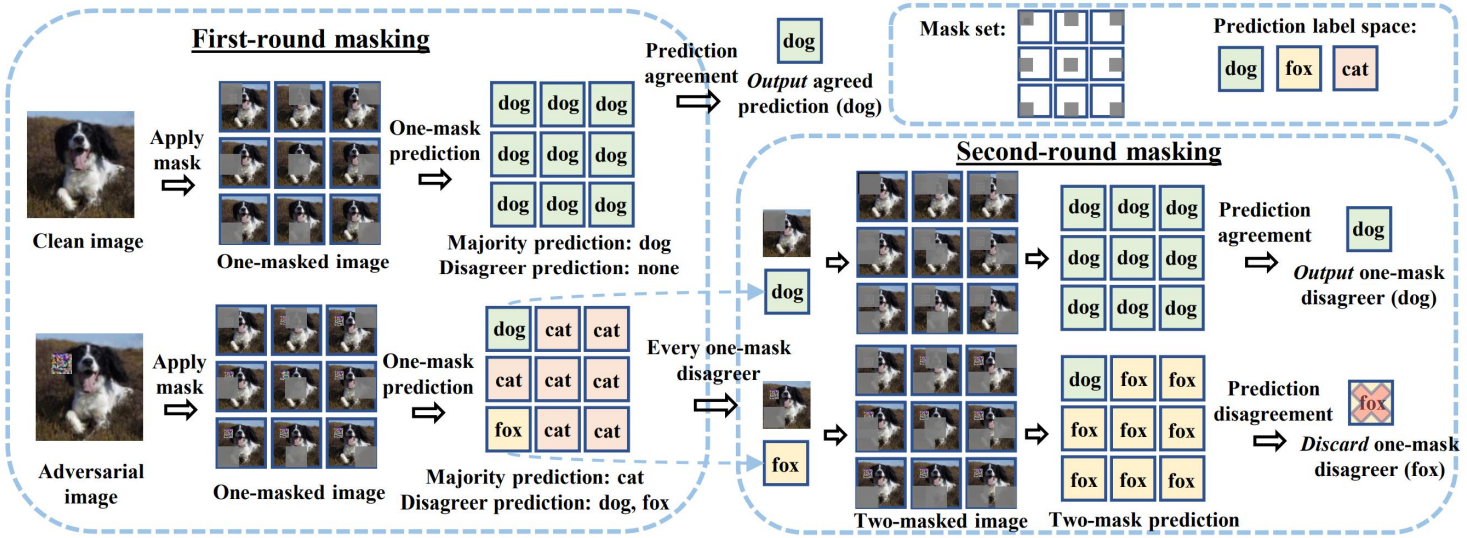
课题四、自适应的对抗图块防御方法

多样化对抗图块攻击下防御的局限

攻击：
成功通常只需要**一个有效**的攻击点

防御：
必须全面无缝，**不能有短板**

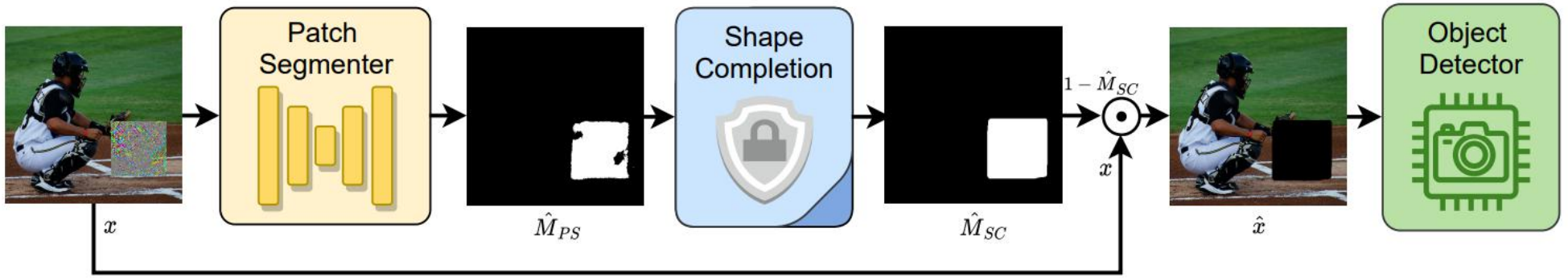
动机：自适应防御多样化的对抗图块攻击



判断-屏蔽图块，PatchCleanser(PC) [Chong,et al.(2022)]

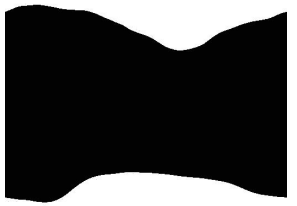
PC：只能用于防御小块的且不能是多区域多块的对抗图块。

SAC：不能很好检测异形的对抗图块，如眼镜形状，且不能是多区域多块的对抗图块，或需要指定块数。



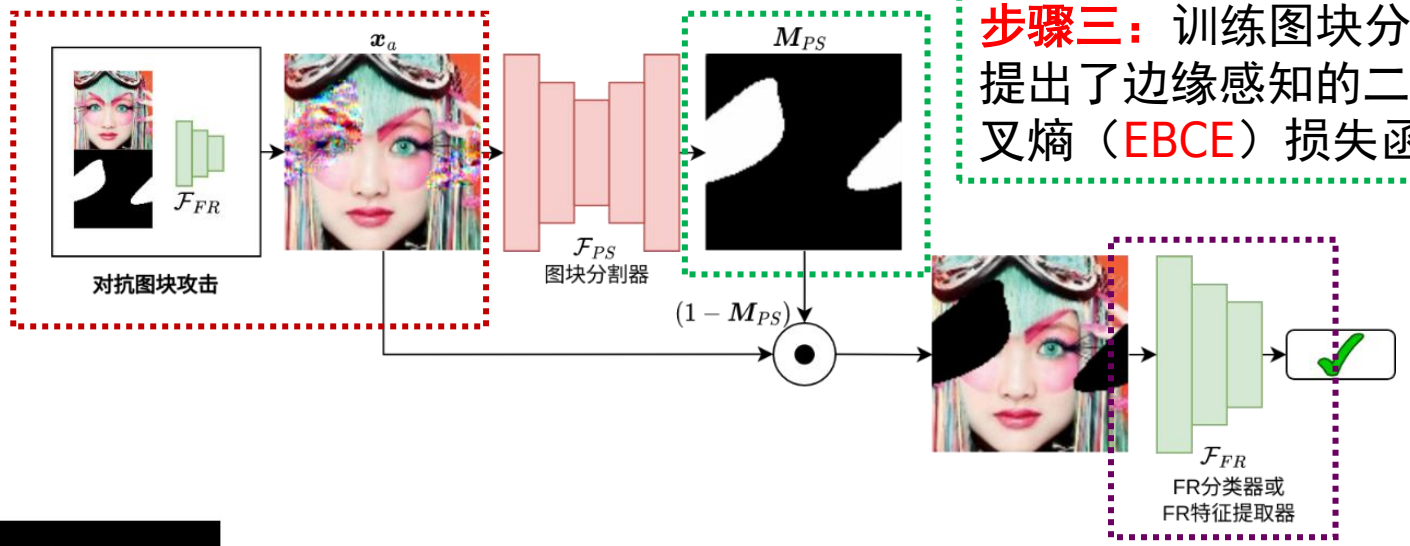
检测-屏蔽图块，Segment and Complete(SAC) [Liu,et al.(2022)]

核心思想：使用图块分割器精确定位并屏蔽对抗图块，以防御其攻击。



基于傅里叶空间采样的掩码，F-mask

步骤二：生成训练数据，提出了基于傅里叶空间采样的对抗图块（F-patch）训练数据。



步骤三：训练图块分割器，提出了边缘感知的二值交叉熵（EBCE）损失函数。

步骤一：增强FR模型遮挡鲁棒性，提出了基于傅立叶空间采样的Cutout（FCutout）数据增强方法。

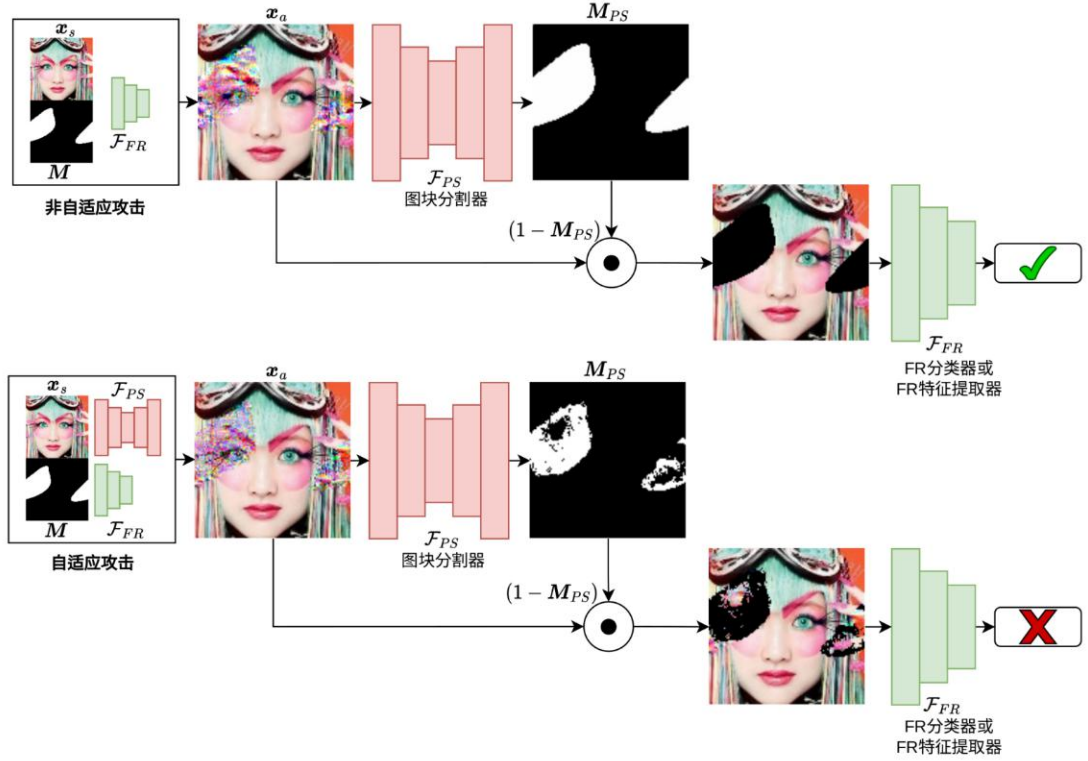
(a) M (b) ∇_M

$\nabla \boldsymbol{M}$的一个示例（论文图6-4）

- 边缘感知的二值交叉熵（EBCE）：

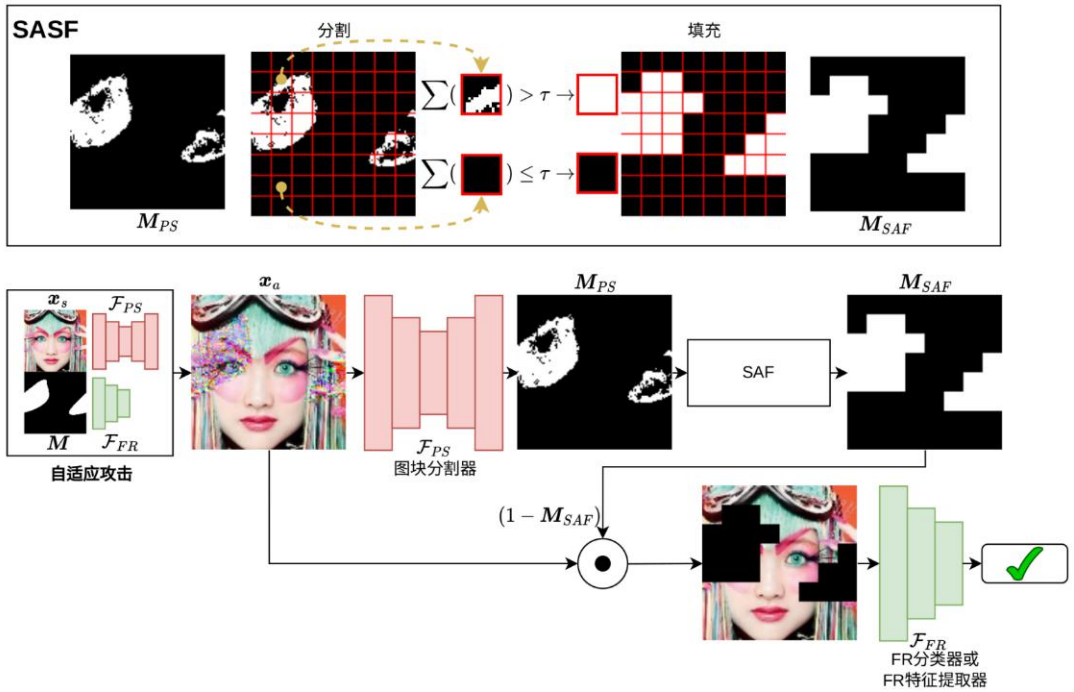
$$\mathcal{L}_{EBCE}(\hat{\boldsymbol{M}}, \boldsymbol{M}) = -\frac{1}{HW}\sum_{i}^{H}\sum_{j}^{W} e^{\beta(\nabla_{\boldsymbol{M}})_{ij}} \left(\boldsymbol{M}_{ij}\log\hat{\boldsymbol{M}}_{ij} + (1-\boldsymbol{M}_{ij})\log(1-\hat{\boldsymbol{M}}_{ij})\right),$$

边缘感知



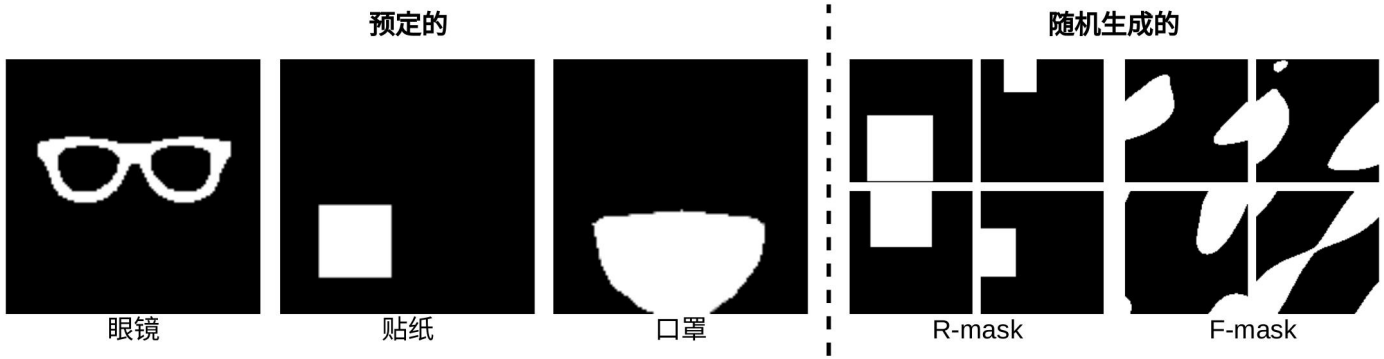
无法防御图块分割器泄漏后的自适应攻击

步骤四：应用自适应分割与填充（SASF）策略。



防御多样的对抗图块攻击

图块分类	训练时	测试
图块掩码	F-mask	F-mask、眼镜、贴纸、口罩、R-mask
攻击者目的	躲避	躲避、冒充
攻击方法	PGDAP	PGDAP、Auto-PGDAP、TAP、EAP、NeRFTAP
目标模型	MobileFaceNet	MobileFaceNet、IResNet50、EfficientNet-B0、Swin-T
开/闭集	闭集	闭集、开集



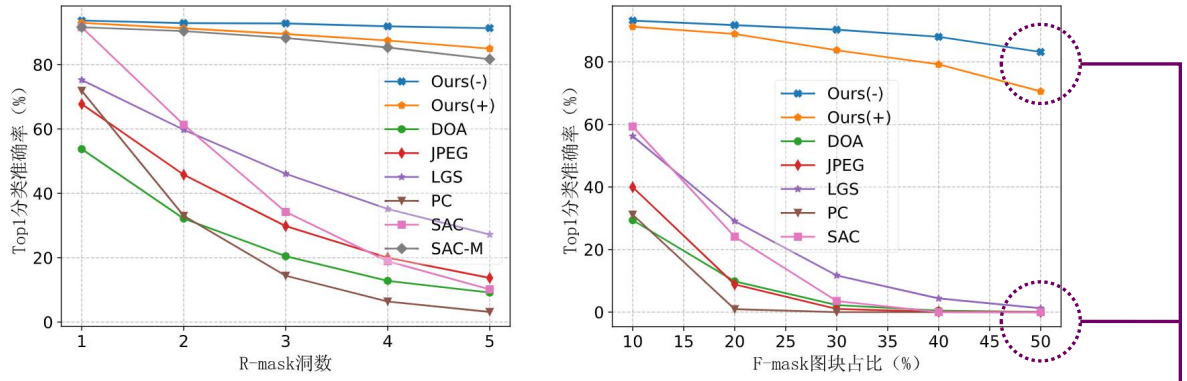
测试时用的攻击图块掩码示例（论文图6-7）

开集人脸识别系统的目标模型（论文表6-3）

目标模型	神经网络结构	损失函数
ArcMob	MobileFaceNet[26]	ArcFace[25]
MagMob	MobileFaceNet[26]	MagFace[99]
MvMob	MobileFaceNet[26]	MV Softmax[100]
MvRes50	IResNet50[23]	MV Softmax[100]
MvSwinT	Swin-T[83]	MV Softmax[100]

闭集，针对未见过的图块掩码和未见过的攻击者目的，各种防御方法的比较结果（论文表6-4）

攻击者目的	防御方法	Clean	掩码 眼镜	贴纸	口罩	R-mask	F-mask
躲避	Cutout[123]	92.97	0.00	0.05	0.00	13.85	5.74
	DOA[68]	80.07	3.15	24.58	1.20	30.77	12.47
	JPEG[38]	88.57	29.96	32.44	0.02	39.90	14.19
	LGS[37]	87.59	47.26	60.75	13.35	54.90	22.40
	PC[70]	92.70	5.18	63.21	0.44	49.30	12.10
	SAC[71]	92.97	30.76	90.66	79.24	88.98	21.83
	AugRmixAT-8-1	93.38	1.90	15.32	0.34	18.23	10.45
	AugRmixAT-8-32	77.44	8.94	25.35	5.64	37.93	18.45
	Ours(−)	94.00	**91.39**	**92.55**	**87.46**	**92.11**	**78.86**
	Ours(+)	**94.01**	84.72	91.28	83.55	89.97	69.79
冒充	Cutout[123]	-	1.01	10.12	0.05	21.32	5.21
	DOA[68]	-	29.21	63.73	16.35	50.15	22.59
	JPEG[38]	-	57.00	56.74	11.43	52.15	24.39
	LGS[37]	-	68.04	77.74	43.72	72.24	39.44
	PC[70]	-	30.69	68.10	20.55	63.52	17.85
	SAC[71]	-	61.99	90.82	80.14	89.53	30.50
	AugRmixAT-8-1	-	10.23	20.85	10.18	32.32	16.34
	AugRmixAT-8-32	-	32.89	68.18	25.23	51.48	26.86
	Ours(−)	-	**91.42**	**92.55**	**87.46**	**92.23**	**78.79**
	Ours(+)	-	84.89	91.28	84.23	90.24	70.56



在不同掩码不同参数下各种防御方法比较结果（论文图6-8）

2、与需要预设孔洞数量的SAC-M不同，RADAP能够自适应地防御未知数量孔洞的攻击。

3、遮挡50%，RADAP性能还是可以稳定在70%以上准确度。

1、特别对于“眼镜”和“F-mask”形状不规则的对抗图块，相比于其他方法，RADAP防御性能更加突出。

针对未见过的目标模型，各种防御方法的比较结果（论文表6-5）

目标模型	防御方法	Clean	掩码				
			眼镜	贴纸	口罩	R-mask	F-mask
IResNet50[23]	JPEG[38]	90.75	24.47	32.27	0.24	40.93	14.55
	LGS[37]	91.46	55.04	67.66	18.63	64.01	28.23
	PC[70]	94.52	2.08	73.89	0.64	56.88	14.14
	SAC[71]	95.12	24.75	93.59	87.25	92.04	24.81
	Ours(−)	**95.13**	**93.87**	**93.89**	**89.95**	**93.73**	**83.78**
	Ours(+)	94.99	89.54	92.87	86.77	92.14	74.62
EfficientNet-B0[84]	JPEG[38]	90.80	34.80	50.36	5.29	52.38	21.99
	LGS[37]	86.16	59.82	74.89	48.23	73.30	41.40
	PC[70]	93.83	0.37	68.04	0.41	51.12	12.26
	SAC[71]	94.39	11.61	92.81	83.86	90.34	21.81
	Ours(−)	**94.48**	**92.27**	**92.96**	**88.51**	**92.48**	**81.12**
	Ours(+)	94.22	86.05	91.77	84.86	90.50	72.26
Swin-T[83]	JPEG[38]	82.79	43.54	61.53	6.44	51.11	20.63
	LGS[37]	89.89	42.95	65.51	10.15	54.55	21.34
	PC[70]	92.14	0.43	73.50	0.00	37.68	10.26
	SAC[71]	92.38	18.94	91.28	82.16	89.04	20.99
	Ours(−)	**92.41**	**91.25**	**91.51**	**87.06**	**91.07**	**80.88**
	Ours(+)	91.37	87.01	90.10	81.60	89.25	72.66

针对未见过的攻击方法，各种防御方法的比较结果（论文表6-6）

攻击方法	防御方法	Clean	掩码				
			眼镜	贴纸	口罩	R-mask	F-mask
APGDAP[11]	DOA[68]	80.07	12.42	36.66	15.18	37.76	20.61
	JPEG[38]	88.57	38.18	47.28	8.07	46.10	19.28
	LGS[37]	87.59	71.46	79.70	52.34	74.86	42.29
	PC[70]	92.70	28.95	66.92	9.06	54.51	19.60
	SAC[71]	92.76	42.82	91.13	79.14	89.38	29.04
	Ours(−)	94.00	**91.66**	**92.90**	**88.49**	**92.36**	**79.87**
	Ours(+)	**94.01**	85.69	91.91	84.51	90.73	71.12
TAP[16]	DOA[68]	80.14	4.00	31.24	2.36	33.16	13.94
	JPEG[38]	88.43	6.99	13.27	0.00	28.59	9.31
	LGS[37]	87.59	45.84	47.72	1.88	43.22	15.88
	PC[70]	92.68	12.02	63.72	0.67	50.88	13.74
	SAC[71]	92.12	42.40	90.30	78.93	89.08	24.77
	Ours(−)	94.00	**89.96**	**92.46**	**87.40**	**92.05**	**78.13**
	Ours(+)	**94.01**	84.61	91.31	82.99	90.01	69.39
EAP	DOA[68]	80.14	3.14	30.31	4.22	27.12	12.28
	JPEG[38]	88.43	8.32	12.62	0.00	19.34	10.21
	LGS[37]	87.59	46.76	45.67	3.23	38.34	14.12
	PC[70]	92.68	11.23	58.67	1.02	45.67	11.45
	SAC[71]	92.12	40.12	90.35	76.34	86.54	20.93
	Ours(−)	94.00	**88.98**	**91.12**	**86.86**	**91.76**	**76.34**
	Ours(+)	**94.01**	83.56	90.46	81.23	90.03	69.34

1. 针对未见过的目标模型与更强的攻击方法时，我们方法同样表现出最好的防御性能。

泛化到开集人脸识别系统上的防御躲避攻击的比较结果（论文表6-7）

目标模型	防御方法	Clean	掩码				
			眼镜	贴纸	口罩	R-mask	F-mask
ArcMob	JPEG[38]	91.80	8.36	42.30	0.10	36.63	22.43
	LGS[37]	88.80	32.83	89.36	50.40	83.86	79.50
	SAC[71]	98.16	0.00	96.96	37.83	82.63	0.00
	Ours(−)	**98.20**	**83.93**	97.06	66.30	**92.40**	**90.16**
	Ours(+)	98.03	73.10	**97.10**	**66.60**	89.23	86.96
MagMob	JPEG[38]	89.63	8.70	43.80	0.23	34.03	25.36
	LGS[37]	91.33	29.43	90.53	57.03	83.50	79.90
	SAC[71]	**98.33**	0.66	97.50	56.06	85.93	0.00
	Ours(−)	98.30	**79.63**	**97.83**	**72.10**	**91.43**	**89.36**
	Ours(+)	98.10	76.03	97.60	66.90	89.30	87.20
MvMob	JPEG[38]	90.80	6.56	38.40	0.13	34.20	25.36
	LGS[37]	91.13	42.16	87.66	57.53	83.53	81.40
	SAC[71]	**98.33**	0.33	**97.23**	56.33	86.30	0.00
	Ours(−)	98.30	**85.13**	97.06	**70.10**	**92.80**	**89.86**
	Ours(+)	98.03	84.13	97.16	68.30	91.70	88.20
MvRes50	JPEG[38]	97.90	49.06	74.10	1.96	52.70	44.26
	LGS[37]	97.63	73.66	96.46	74.26	91.96	91.96
	SAC[71]	**99.53**	6.46	99.23	81.80	94.60	0.20
	Ours(−)	**99.53**	**97.93**	**99.30**	**92.80**	**97.53**	**97.50**
	Ours(+)	**99.53**	96.46	99.23	90.60	96.50	96.10
MvSwinT	JPEG[38]	98.20	91.00	94.46	76.73	90.86	91.96
	LGS[37]	99.30	97.10	99.46	92.86	96.96	97.03
	SAC[71]	99.60	2.23	**99.60**	85.76	94.60	0.00
	Ours(−)	**99.66**	**98.83**	99.50	**95.56**	**97.60**	**97.96**
	Ours(+)	99.53	97.73	99.46	93.20	96.70	96.36

开集人脸识别系统上，防御EAP和NeRFTAP冒充攻击的比较结果（论文表6-8）

防御方法	EAP		NeRFTAP	
	数据集		掩码	
	LFW	CelebA	眼部 & 鼻子	口罩
UN	0.4647+-0.0888	0.4453+-0.0919	0.3461+-0.0856	0.2562+-0.0813
JPEG[38]	0.4702+-0.0853	0.4389+-0.0893	0.3308+-0.0822	0.2447+-0.0810
LGS[37]	0.3573+-0.1100	0.3354+-0.1027	0.1533+-0.0982	0.1942+-0.0867
SAC[71]	0.4354+-0.1064	0.4335+-0.0951	0.3004+-0.0877	0.2300+-0.0980
Ours(−)	**0.1628+-0.1147**	**0.1240+-0.1028**	**0.1457+-0.0838**	**0.1413+-0.0965**
Ours(+)	0.2979+-0.1101	0.2483+-0.1117	0.1903+-0.0754	0.1766+-0.1037

1、直接泛化到开集人脸识别系统，同样针对于不同的攻击目的，不同的目标模型，不同的攻击方法，RADAP基本上还是保持着最优防御性能。

针对FCutout的消融实验的结果（论文表6-9）

防御方法		Clean	掩码 眼镜	贴纸	口罩	R-mask	F-mask
UN	Vanilla	89.43	0.00	0.04	0.00	14.12	5.51
	Cutout[123]	92.97	0.00	0.05	0.00	13.85	5.71
	FCutout	**94.09**	0.00	**0.32**	0.00	**17.22**	**6.40**
GT	Vanilla	89.43	64.23	82.96	61.95	79.27	55.32
	Cutout[123]	92.97	88.70	90.83	83.08	90.82	71.93
	FCutout	**94.09**	**91.42**	**92.56**	**87.46**	**92.12**	**78.90**
SAC[71]	Vanilla	89.31	3.69	82.21	48.92	72.20	17.58
	Cutout[123]	92.97	**30.76**	90.66	79.24	88.98	21.83
	FCutout	**94.02**	29.50	**92.49**	**83.47**	**90.17**	**24.09**
Ours(−)	Vanilla	89.43	63.81	82.97	61.94	79.29	55.06
	Cutout[123]	92.97	88.62	90.85	83.08	90.79	71.89
	FCutout	**94.00**	**91.39**	**92.55**	**87.46**	**92.11**	**78.86**
Ours(+)	Vanilla	88.62	48.09	74.96	51.38	73.51	42.64
	Cutout	92.86	**85.14**	89.60	79.78	88.90	65.19
	FCutout	**94.01**	84.72	**91.28**	**83.55**	**89.97**	**69.79**

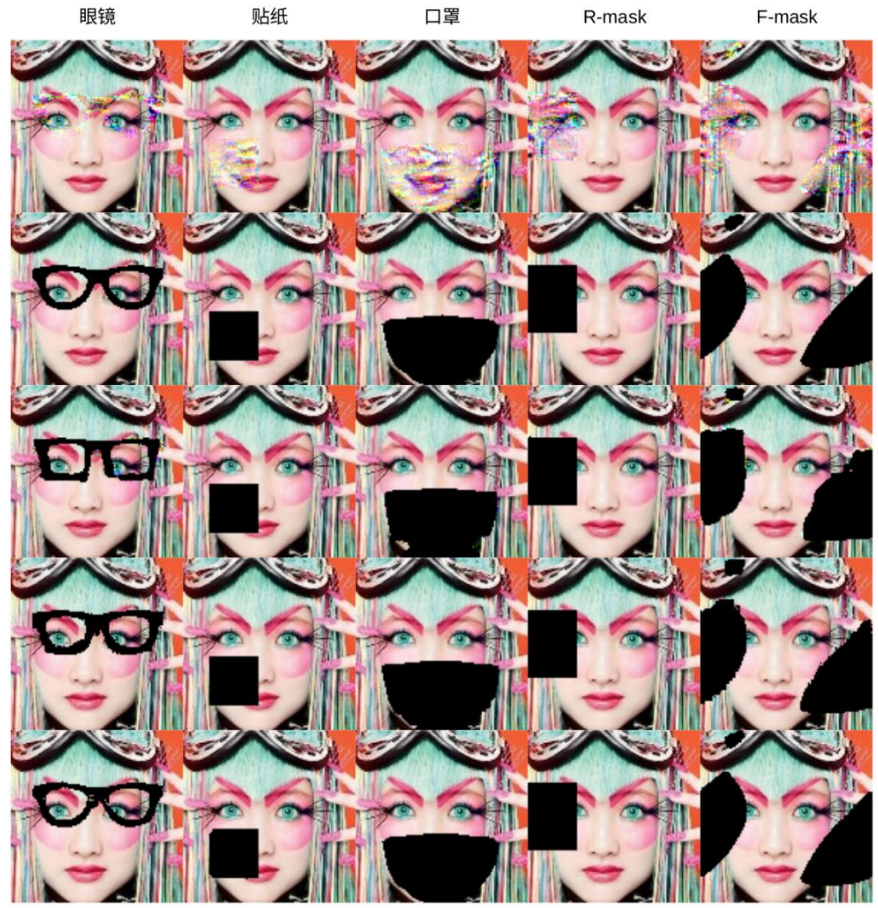
1、FCutout具有更强的遮挡鲁棒性。

针对F-patch和EBCE的消融实验的结果（论文表6-10）

防御方法	Clean	掩码 眼镜	贴纸	口罩	R-mask	F-mask
R-patch + BCE	**94.08**	78.55	92.53	86.47	92.11	65.87
R-patch + EBCE	94.02	**90.64**	**92.53**	**87.25**	**92.12**	**77.57**
F-patch + BCE	93.93	91.20	92.30	87.34	92.05	78.76
F-patch + EBCE	**94.00**	**91.39**	**92.55**	**87.46**	**92.11**	**78.86**

2、F-patch对于R-mask防御性能轻微下降，F-mask下显著提升。

3、『眼镜』越来越精准

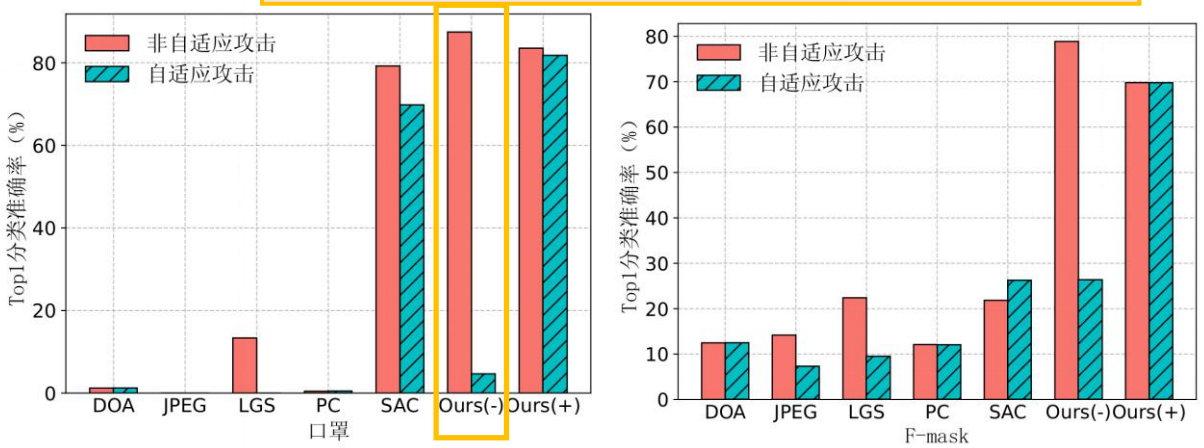


针对F-patch和EBCE的消融实验的可视化结果（论文图6-9）

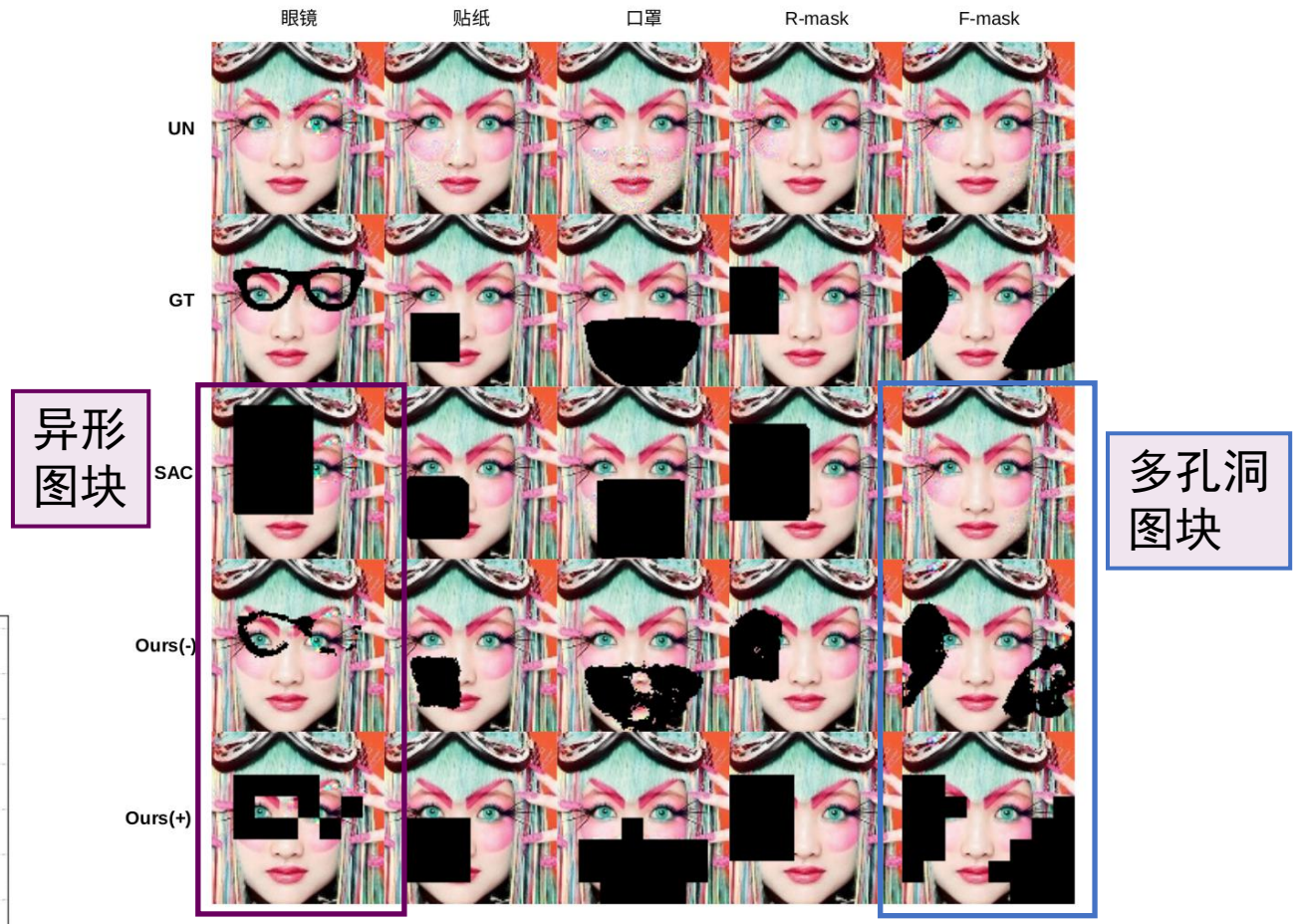
防御方法泄漏后防御结果比较。（论文表6-11）

防御方法	Clean	掩码				
		眼镜	贴纸	口罩	R-mask	F-mask
DOA[68]	80.07	3.15	24.58	1.20	30.77	12.47
JPEG[38]	88.57	1.33	2.94	0.00	21.48	7.29
LGS[37]	87.59	18.09	25.11	0.00	28.47	9.53
PC[70]	92.70	5.18	63.21	0.44	49.30	12.10
SAC[71]	92.97	14.42	88.88	69.83	85.37	26.29
Ours(−)	94.00	16.45	71.90	4.61	61.38	26.40
Ours(+)	**94.01**	**82.85**	**91.15**	**81.81**	**89.76**	**69.75**

1、没有SASF策略无法防御自适应攻击



防御自适应攻击和非自适应攻击的比较结果



防御自适应攻击可视化结果（论文图6-11）

RADAP的核心贡献

- **RADAP防御策略：**提出一种鲁棒且自适应的对抗图块防御方法，有效应对多样化的对抗图块攻击。
- **关键技术创新：**RADAP结合了几项关键技术创新，包括FCutout、F-patch、EBCE损失函数和SASF策略，这些组合提高了防御策略的效果和适应性。
- **实验验证：**通过广泛的实验设置测试，验证了RADAP在多种环境下的有效性，证明了其在实际应用中的潜力。

现有系统的启示

- **安全威胁的实际影响：**对抗性攻击技术的快速发展要求现有系统必须不断更新和改进防御措施，以应对日益复杂的安全威胁。
- **更全面的保护：**防御策略必须广泛覆盖多样的对抗攻击，确保系统即使面对未知个体和新型攻击手法也能维持高度的安全性和准确性。

南京大学
NANJING UNIVERSITY

三、总结与展望

工作总结　未来展望

对抗攻击

课题一：基于二维变换的可迁移对抗图块攻击方法

针对黑盒攻击在人脸识别中的低成功率问题，提出了EAP攻击方法。该方法通过二维变换技术，精确识别系统的脆弱点，有效提高了对抗样本的迁移性和黑盒攻击的成功率。

课题二：基于三维神经辐射场的可迁移对抗图块攻击方法

针对三维环境下人脸识别系统的黑盒攻击困难问题，提出了基于三维神经辐射场技术的NeRFTAP攻击方法，使得对抗性攻击能够扩展到三维空间，显著增强了在复杂三维场景中的攻击适用性和有效性。

对抗防御

课题三：增强模型多重鲁棒性的对抗训练防御方法

针对人脸识别模型对抗鲁棒性和泛化能力的平衡问题，设计了AugRmixAT对抗训练防御框架，该框架结合数据增强与对抗性训练策略，有效促进了人脸识别模型在对抗鲁棒性与泛化能力间的良性平衡。同时提升了模型在对抗、遮挡、腐化等多方面的鲁棒性。

课题四：自适应的对抗图块防御方法

针对现有防御机制难以抵御多样化的对抗图块攻击的问题，提出了RADAP防御策略，专门针对多样化的对抗图块进行自适应防御，进一步加强了人脸识别系统与模型的安全性

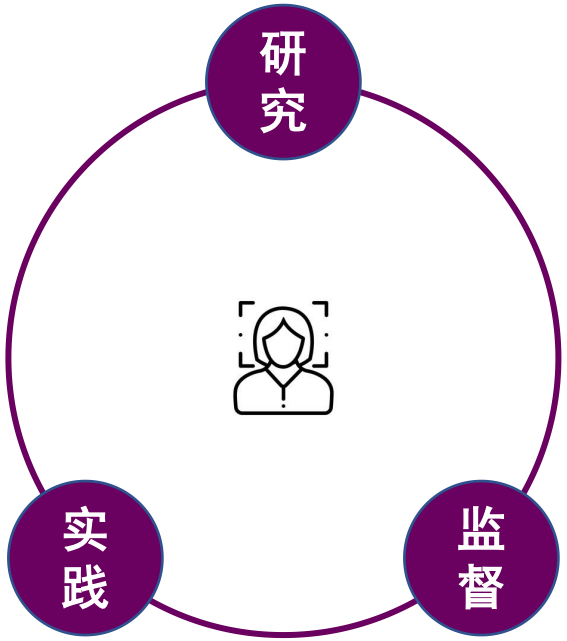
未来展望

更复杂环境下的攻击方法研究

进一步探索更加复杂环境下（如不同光照，天气等）的对抗攻击方法。

更全面的防御策略研究

探索和开发综合性防御策略，以抵御不断演化的对抗性攻击。



赋能实际应用中

将研究成果应用于实际行业中，进一步提升使用人脸识别系统的安全性。

推进伦理和法规制定

结合计算机科学，数据分析和法律伦理，推动制定伦理和法规，确保人脸识别技术的合理与公正使用。

四、科研成果总结

学术论文 其他成果

已发表论文（第一作者，共4篇）

1. **Xiaoliang Liu**, Furao Shen, Jian Zhao and Changhai Nie. AugRmixAT: A data processing and training method for improving multiple robustness and generalization performance [C]. IEEE International Conference on Multimedia and Expo (ICME). 2022: 1-6.（CCF-B类会议，EI索引）
2. **Xiaoliang Liu**, Furao Shen, Jian Zhao and Changhai Nie. Self-supervised learning of monocular 3D geometry understanding with two-and three-view geometric constraints[J]. The Visual Computer, 2024, 40(2): 1193-1204.（CCF-C类期刊，SCI索引，JCR-Q2，中科院3区）
3. **Xiaoliang Liu**, Furao Shen, Jian Zhao and Changhai Nie. EAP: An effective black-box impersonation adversarial patch attack method on face recognition in the physical world[J]. Neurocomputing, 2024: 127517.（CCF-C类期刊，SCI索引，JCR-Q2，中科院2区）
4. **Xiaoliang Liu**, Furao Shen, Jian Zhao and Changhai Nie. RandoMix: a mixed sample data augmentation method with multiple mixed modes[J]. Multimedia Tools and Applications, 2024: 1-17.（CCF-C类期刊，SCI索引，JCR-Q2，中科院4区）

在投论文（第一作者，共2篇）

1. **Xiaoliang Liu**, Furao Shen, Feng Han, Jian Zhao and Changhai Nie. NeRFTAP: Enhancing transferability of adversarial patches on face recognition using neural radiance fields[J]. arXiv preprint arXiv:2311.17332, 2023.（CCF-B类期刊，Under Review）
2. **Xiaoliang Liu**, Furao Shen, Jian Zhao and Changhai Nie. RADAP: A robust and adaptive defense against diverse adversarial patches on face recognition[J]. arXiv preprint arXiv:2311.17339, 2023.（CCF-B类期刊，Under Review）

学术论文 其他成果

发明专利

1. 申富饶，高可攀，**刘小亮**，等. 一种融合UWB和LiDAR的室内定位方法：202011520518[P] [2024-02-28].（核心算法的提出）

已出版的专著

1. 申富饶.自组织增量学习神经网络[M]. 电子工业出版社，2024.（负责第一章的写作）

科研课题

1. 国家电网总部科技项目，基于多维图像智能匹配及识别技术的变电站高清视频和机器人联合巡检技术研究及应用，项目号：520950200009。（参与）
2. 科技部重大项目-科技创新2030项目，基于神经可塑性的脉冲网络高效学习机制与类脑智能系统，项目号：2021ZD0201300。（参与）
3. 国家自然科学基金面上项目，面向增量式无监督学习的新型神经网络研究，项目号：62276127。（参与）

竞赛获奖

1. 2022年08月-2023年03月，首届粤港澳大湾区（黄埔）国际算法算例大赛，深度学习模型的对抗鲁棒防御算法，入围决赛奖。（单人成队，前15名队伍进入决赛）

盲审结果

盲审专家	总分	总体评价	是否同意答辩	熟悉程度
专家一	85	良好	修改后直接答辩	很熟悉
专家二	83	良好	修改后直接答辩	熟悉
专家三	88	良好	修改后直接答辩	很熟悉

南京大学
NANJING UNIVERSITY

谢谢大家
敬请各位专家老师批评指正