ICCV HONOLULU HAWAII
OCT 19-23, 2025

FaceShield: Defending Facial Image against Deepfake Threats

Jaehwan Jeong[1], Sumin In[1], Sieun Kim[1], Hannie Shin[1],
Jongheon Jeong[1], Sangho Yoon[2], Jaewook Chung[3], Sangpil Kim[1,*]

[1] Korea University [2] KAIST [3] Samsung Research

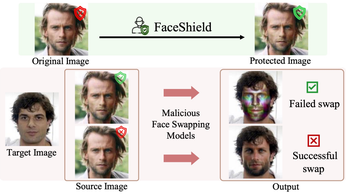
Paper Link:

고려대학교 KOREA UNIVERSITY

KAIST

SAMSUNG Research

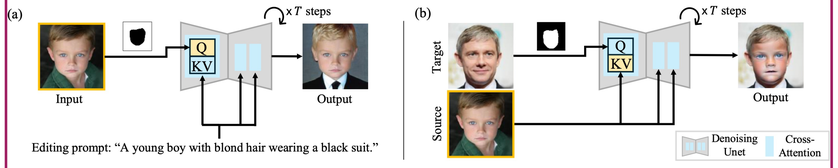
Motivation



- **Deepfake threats** are escalating.
- However, current research remains largely **detection focused**, and the few proactive defenses are mostly from the **GAN era** and are not suitable for diffusion models.
- Moreover, existing diffusion attacks concentrate on **single image editing** rather than the two image conditioning used in deepfakes.

→ A new attack method is needed for **diffusion-based deepfake models**

Problem Settings



(a) Image editing models: Single query image Q with a text prompt (prompt → K, V)

→ The standard assumption in **Prior** diffusion-based adversarial attacks.
Attacks focus on the **Query path.** * **AdvDM, Mist, PhotoGuard, SDST**

(b) Deepfake models: Two images. Target = Q, source provides K,V via cross-attention

→ Requires an adversarial attack on the **Conditioning path.** * **FaceShield (Ours)**

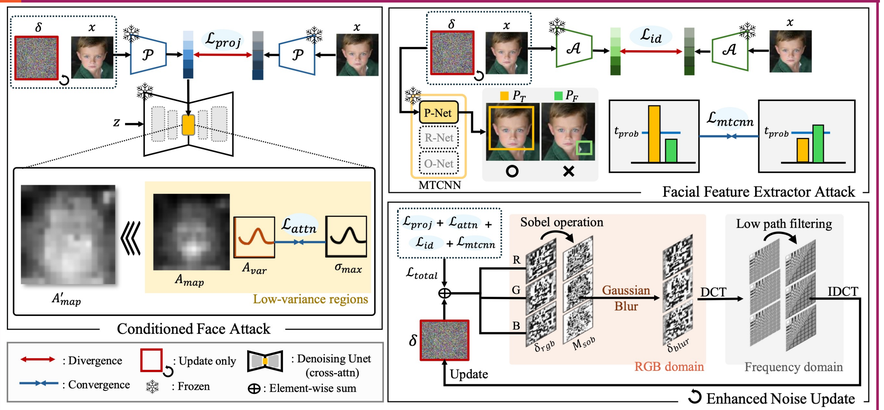
Contributions

i. **We introduce the first adversarial attack tailored to image-conditioned, diffusion-based deepfake models.**

ii. **A Gaussian-blur guided PGD update boosts imperceptibility, and Low-pass filtering makes perturbations robust to JPEG/purification defenses.**

iii. **Extensive experiments show state-of-the-art attack effectiveness with less visible noise than prior diffusion-based methods.**

Overall Pipeline



- **Conditioned Face Attack :** Modulates cross attention variance.
- **Facial Feature Extractor Attack :** Disrupts face recognition in deepfakes.
- **Enhanced Noise Update :** Gaussian blur and low pass for compression robustness.

Algorithm

Algorithm 1: FaceShield

Input: image x, steps N, noise clamp ϵ, step size α, MTCNN detection threshold t_{prob}, threshold weight β, CLIP Image Projector $\mathcal{P}$, Mid-layer cross-attention variance in Stable Diffusion A'_{var}, MTCNN P-Network $\mathcal{T}$, ArcFace $\mathcal{A}$

Result: protected image x_{adv}

Initialize adversarial perturbation $\delta \leftarrow 0$, and protected image $x_{\text{adv}} \leftarrow x$

for $n = 1, \ldots, N$ **do**

  $\mathcal{L}_{\text{proj}} \leftarrow \|\mathcal{P}(x+\delta) - \mathcal{P}(x)\|_1$

  $\mathcal{L}_{\text{attn}} \leftarrow \|(\sigma_{\max} - A'_{\text{var}}) \odot M_{\text{var}}\|_2$, where $\sigma_{\max}$ derived from Eq. (8), and M_{var} from Eq. (6)

  $\mathcal{L}_{\text{mtcnn}} \leftarrow \|(\mathcal{T}(x+\delta) - p_{\text{gt}}) \odot M_{\text{prob}}\|_2$, where $p_{\text{gt}} = [t_{\text{prob}} + \beta t_{\text{prob}} - \beta]^T$, and M_{prob} from Eq. (9)

  $\mathcal{L}_{\text{id}} \leftarrow 1 - \frac{\mathcal{A}(x+\delta)\cdot\mathcal{A}(x)}{\|\mathcal{A}(x+\delta)\|_2 \|\mathcal{A}(x)\|_2}$

  Compute the total attack loss:
  $\mathcal{L}_{\text{total}} = \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{mtcnn}}\mathcal{L}_{\text{mtcnn}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}}$

  Update adversarial perturbation:
  $\delta \leftarrow \alpha \cdot \text{sign}(\nabla_{x_{\text{adv}}}\mathcal{L}_{\text{total}})$

  $\delta_{\text{blur}} \leftarrow \textbf{GaussianBlur}(\delta)$

  $\delta'_{\text{rgb}} \leftarrow \textbf{LowPassFilter}(\delta_{\text{blur}})$

  $x_{\text{adv}} \leftarrow x_{\text{adv}} - \delta'_{\text{rgb}}$

  $x_{\text{adv}} \leftarrow x + \text{clip}(x_{\text{adv}} - x, -\epsilon, \epsilon)$

end

Clip the image range: $x_{\text{adv}} \leftarrow \text{clip}(x_{\text{adv}}, 0, 255)$

Algorithm 2: Adversarial loss in cross attention.

Input: perturbation δ, query embedding Q_x, original source face embedding K_x, adversarial source face embedding $K_{(x+\delta)}$, low variance threshold t_{var}, maximum variance value $\sigma_{\max}$, low variance mask M_{var}, attention loss $\mathcal{L}_{\text{attn}}$, attention loss function $\mathcal{F}$

Result: stored low-variance mask M_{var}, added attention loss $\mathcal{L}_{\text{attn}}$

if M_{var} *is not precomputed* **then**

  // Construct Ground Truth

  Compute original attention map: $A_{\text{map}} \leftarrow \textbf{Softmax}(Q_x K_x^T/\sqrt{d})$

  Compute variance: $A_{\text{var}} \leftarrow \textbf{Var}(A_{\text{map}})$

  Calculate low-variance threshold: $P_{t_{\text{var}}} \leftarrow \textbf{Quantile}(A_{\text{var}}, t_{\text{var}})$

  Generate low-variance mask: $M_{\text{var}} \leftarrow \textbf{Mask}(A_{\text{var}}, P_{t_{\text{var}}})$

  Store M_{var} for applying adversarial noise

end

else

  // Compute Adversarial Loss

  Compute adversarial attention map: $A'_{\text{map}} \leftarrow \textbf{Softmax}(Q_x K_{(x+\delta)}^T/\sqrt{d})$

  Compute variance: $A'_{\text{var}} \leftarrow \textbf{Var}(A'_{\text{map}})$

  Calculate attention loss in low-variance regions: $\mathcal{L}_{\text{attn}} \leftarrow \mathcal{L}_{\text{attn}} + \mathcal{F}(\Delta)$, where $\Delta = (\sigma_{\max} - A'_{\text{var}}) \odot M_{\text{var}}$

end

Subsequent steps are not shown here.

▶ Attention Disruption Attack (Core)

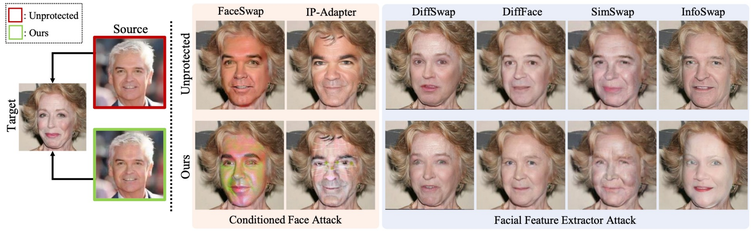
$$A_{\text{var}} = \frac{1}{\text{seq}} \sum_{i=1}^{\text{seq}} \left(A_{\text{map}}[:,:,i] - \bar{A}_{\text{map}}\right)^2 \in \mathbb{R}^{h \times \text{res}}$$

$$\mathcal{L}_{\text{attn}}(\delta;\ x,\ \sigma_{\max}) = \|(\sigma_{\max} - A'_{\text{var}}) \odot M_{\text{var}}\|_2$$

- $A_{\text{map}} \in \mathbb{R}^{h \times \text{res} \times \text{seq}}$: Attention map
- $\bar{A}_{\text{map}}$: Mean attention map
- $\sigma_{\max}$: Maximum variance
- M_{var}: Low variance mask

Results

▶ Qualitative Demonstration of Generality



▶ Quantitative Comparison of Method Performance

Model	DiffFace [25]				DiffSwap [68]				FaceSwap [50]				IP-Adapter [61]			
Dataset	CelebA-HQ [23]															
Method	$L_2\uparrow$	ISM↓	PSNR↓	HE↑	$L_2\uparrow$	ISM↓	PSNR↓	HE↑	$L_2\uparrow$	ISM↓	PSNR↓	HE↑	$L_2\uparrow$	ISM↓	PSNR↓	HE↑
AdvDM [32]	0.021	0.471	39.368	4.22	0.068	0.199	28.362	4.68	0.303	0.245	21.615	4.52	0.207	0.235	25.332	2.76
Mist [31]	0.021	0.468	39.443	3.94	0.067	0.201	28.384	4.18	0.287	0.230	22.263	4.78	0.152	0.265	28.213	4.26
PhotoGuard [42]	0.022	0.469	39.194	3.82	0.068	0.201	28.292	4.58	0.282	0.238	22.316	4.44	0.153	0.268	28.101	4.44
SDST [60]	0.021	0.470	39.512	4.08	0.067	0.207	28.383	5.04	0.274	0.261	22.582	4.68	0.147	0.273	28.440	4.32
Ours	**0.044**	**0.243**	**32.052**	**5.76**	**0.072**	**0.163**	**27.833**	**6.20**	**0.336**	**0.194**	**20.759**	**6.16**	**0.350**	**0.072**	**20.266**	**6.60**
Ours (Q=75)	0.043	0.259	32.259	-	0.070	0.169	28.034	-	0.317	0.209	21.286	-	0.326	0.112	20.867	-
Dataset	VGGFace2-HQ [5]															
Method	$L_2\uparrow$	ISM↓	PSNR↓	HE↑	$L_2\uparrow$	ISM↓	PSNR↓	HE↑	$L_2\uparrow$	ISM↓	PSNR↓	HE↑	$L_2\uparrow$	ISM↓	PSNR↓	HE↑
AdvDM [32]	0.042	0.479	33.064	3.68	0.105	0.215	24.769	4.78	0.419	0.361	18.596	4.38	0.251	0.271	23.250	2.36
Mist [31]	0.041	0.478	33.215	4.26	0.102	0.227	24.964	3.94	0.379	0.259	19.626	4.50	0.181	0.291	26.070	4.10
PhotoGuard [42]	0.043	0.479	32.938	3.96	0.110	0.215	24.272	4.18	0.373	0.266	19.655	4.14	0.180	0.294	26.157	3.82
SDST [60]	0.041	0.483	33.242	5.30	0.107	0.225	24.506	4.58	0.359	0.258	19.996	4.14	0.166	0.292	26.784	4.06
Ours	**0.062**	**0.278**	**29.204**	**6.10**	**0.113**	**0.177**	**24.054**	**6.12**	**0.453**	**0.237**	**17.919**	**6.16**	**0.382**	**0.112**	**19.478**	**6.42**
Ours (Q=75)	0.060	0.308	29.435	-	0.112	0.185	24.201	-	0.421	0.237	18.573	-	0.377	0.167	19.618	-

▶ Perturbation Imperceptibility

Dataset	CelebA-HQ [23]				
Method	LPIPS↓	PSNR↑	SSIM↑	FR↑	HE↑
AdvDM [32]	0.4214	30.4476	0.8438	2.1077	3.86
Mist [31]	0.5492	29.9935	0.8684	1.6583	4.70
PhotoGuard [42]	0.5515	29.9127	0.8669	1.6538	4.82
SDST [60]	0.5409	31.4762	0.9033	1.6767	5.12
Ours	**0.2017**	**32.6289**	**0.9394**	**18.4651**	**5.64**
Dataset	VGGFace2-HQ [5]				
Method	LPIPS↓	PSNR↑	SSIM↑	FR↑	HE↑
AdvDM [32]	0.4108	30.2523	0.8436	2.0667	3.66
Mist [31]	0.5208	29.9068	0.8721	1.6872	4.34
PhotoGuard [42]	0.5221	29.8204	0.8712	1.6824	4.62
SDST [60]	0.5060	31.3545	0.9092	1.6892	4.48
Ours	**0.1941**	**31.5799**	**0.9341**	**18.0400**	**5.28**

▶ Purification Robustness