FaceOff: Preventing Unauthorized Text-to-Image Identity Customization

Mingwang Hu[1], Chenyu Zhang[1], Zili Yi[2], Lanjun Wang[1]*
[1] School of New Media and Communication, Tianjin University
[2] School of Intelligence Science and Technology, Nanjing University

***Abstract*—Recently, tuning-free techniques such as PhotoMaker have advanced text-to-image identity (ID) customization. Unlike tuning-based approaches, these techniques employ powerful facial encoders to extract ID information from a single portrait, enabling efficient customization in a single inference pass. However, their misuse can exacerbate the generation of misleading and harmful content, endangering individuals and society. To mitigate this risk, existing methods introduce protective perturbations into user portrait images to distort ID information in the customized images. We identify two limitations of these methods: 1) copyright infringement risk caused by identified faces in customized images; 2) lack of robustness against image transformations. To address these limitations, we propose FaceOff, a framework that protects portrait images from the threat of tuning-free ID customization by disturbing the ensemble of image encoders. Specifically, to reduce the recognized faces in customized images, we design a contrastive loss to shift the image semantics to a target ID and deviate the image semantics away from the original ID. In addition, we introduce a Gaussian augmentation module to mitigate the protection degradation under image transformations. FaceOff outperforms the SOTA method by 12.65% of FDFR and 4.70% of ISM across three customization models and two facial benchmarks on average. The code is available at https://github.com/huzimun/FaceOff.**

***Index Terms*—DeepFake Defense, Adversarial Examples, Diffusion Model, Identity Customization**

I. Introduction

Due to the powerful generative capabilities of diffusion models [1]–[3], many researchers try to explore identity (ID) customization based on them [4], [5]. Given a **text prompt** and **reference images** with the same ID, users utilize text-to-image ID customization models to generate **customized images** of the input ID. Current text-to-image ID customization models can be divided into two types: tuning-based and tuning-free. Tuning-based models [4]–[7], such as Textual Inversion [4] and DreamBooth [5], need to first learn the ID information of reference images by a fine-tuning process, which costs a lot of time and computational resources for each customization task, hindering their real-world applicability. To ease resource demand, various tuning-free models [8]–[12] have emerged. These models utilize facial encoders to encode reference images with ID information into the generation process, achieving better fidelity and efficiency. However, with such powerful tools, attackers can use some publicly posted photos with a person's face to generate harmful content.

To mitigate this risk, some studies [13]–[20] propose creating **protected images** by adding perturbations to reference images, thereby degrading the fidelity of customized images. However, most approaches [13]–[19] focus on tuning-based models, relying on fine-tuning to incorporate perturbations into the learning and generation of ID information. This dependence constrains the effectiveness of these methods concerning tuning-free models, which directly encode the reference image without undergoing a fine-tuning process. In addition, their protection efficacy under image transformations degrades significantly [21], [22]. A latest work IDProtector [20] proposes a defense against tuning-free models. However, customized images protected by IDProtector still contain human faces that might infringe copyrighted portrait images.

This work aims to provide universal and robust protection against tuning-free ID customization, which presents two main challenges: **1. Universality.** Since attackers can utilize various tuning-free models, protection needs to work on different customization models rather than a specific model. **2. Robustness.** As attackers can apply various image transformations to remove protective perturbations, protection must withstand both common image transformations, such as JPEG compression, and advanced adversarial purification methods [21], [22].

To address these challenges, we propose FaceOff, a framework for defending against tuning-free customization models. Firstly, to handle diverse tuning-free models, we disrupt the image encoders of three representative methods, including IP-Adapter [8], IP-Adapter Plus [8], and PhotoMaker [10], to deviate protected image semantics from the original. However, as these models are tailored for portrait generation, customized images may still contain recognizable faces, posing copyright risks. To mitigate this, we introduce a contrastive loss that pulls protected semantics toward a target image while pushing them away from the original, encouraging the protected customized images to exhibit target-like characteristics with fewer recognizable faces. Secondly, to improve robustness against transformations, we introduce a Gaussian augmentation loss by incorporating Gaussian filtering into the contrastive loss. While this enhances protection under transformations, it degrades performance in standard customization

This work was supported by the National Natural Science Foundation of China under Grants 62572346 and 62202329, the Jiangsu Provincial Science and Technology Major Project under Grant No. BG2024042, and the Shen Zhen Key Laboratory of Media Security under Grant No. SYSPG20241211174032004. * Corresponding author: Lanjun Wang (wanglanjun@tju.edu.cn).

due to overfitting to the filtering. To address this, we combine the contrastive and Gaussian augmentation losses to achieve robustness to transformations while preserving performance in standard customization. Our main contributions are as follows:

- We propose FaceOff, a framework that protects portrait images against tuning-free ID customization.
- We design a joint loss, consisting of a contrastive loss and a Gaussian augmentation loss, to reduce the identified human faces in the customized images and enhance the protection robustness against image transformations.
- Experiments show that FaceOff outperforms the SOTA method by 12.65% of FDFR and 4.70% of ISM across three models and two benchmarks on average.

II. Related Works

A. Text-to-Image ID Customization

Text-to-image ID customization methods utilize pre-trained diffusion models to generate images of specific IDs with several reference images and text descriptions. Current text-to-image ID customization models can be divided into two categories [10]: tuning-based and tuning-free. Tuning-based models [4]–[7], [23], [24] usually rely on additional optimization during the test phase. Although tuning-based models have achieved commendable results, customizing each ID requires substantial time for fine-tuning, making the process economically expensive and limiting its applications. To reduce the resource-consuming nature of customization, many tuning-free models [8]–[12], [25]–[29] have emerged. These models forgo fine-tuning for each ID, instead, they directly encode ID information into the generation process via pre-training an ID adapter. They typically utilize an encoder (e.g., a CLIP image encoder [30]) to extract the ID feature. The extracted feature is integrated into the base diffusion model in a specific way (e.g., embedded into cross-attention layers). Due to their high fidelity and efficiency, tuning-free models have rapidly gained popularity and are widely adopted for customization. In this work, we focus on defending against these emerging tuning-free customization models.

B. Protection against Unauthorized ID Customization

With the application of powerful ID customization tools, the public images of individuals are at risk of being exploited for malicious use. To defend against unauthorized ID customization, existing methods [13]–[19], [31] inject adversarial perturbations into portrait images to degrade both the ID consistency and visual quality of the generated customized images. However, these methods mainly focus on tuning-based models, and their protection performance degrades substantially on tuning-free models [20]. To fill this gap, we propose to defend against tuning-free ID customization models based on contrastive loss and Gaussian augmentation. Concurrently, Song et al. [20] introduce IDProtector, which injects imperceptible perturbations into portrait images via a single forward pass. The customized images protected by IDProtector still contain many recognizable faces, which may pose potential copyright infringement risks. In contrast, our contrastive-loss-based defense effectively reduces the number of recognizable faces in protected customized images, thereby mitigating this risk.

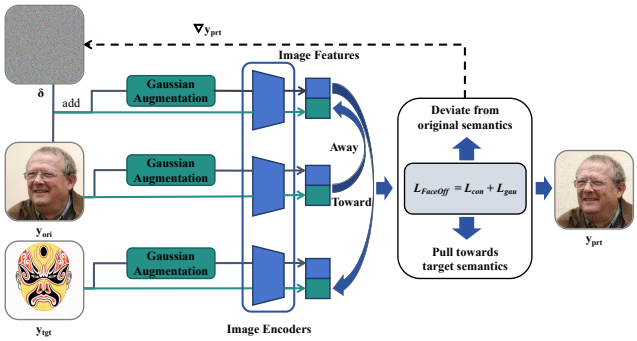

Fig. 1. Framework of FaceOff. FaceOff generates a protected image by injecting learnable perturbations into the original image, applies Gaussian filtering to enhance robustness against transformations, and optimizes contrastive and Gaussian augmentation losses in the semantic feature space.

III. Problem Statement

We describe our task as follows. A user (protector) wants to protect his image y_{ori} with ID ori from unauthorized use, i.e., to generate customized images using a tuning-free ID customization model M. To achieve this, the protector injects an adversarial perturbation δ into the original image y_{ori} to create a protected image, denoted y_{prt} (i.e., $y_{ori}+\delta$), which is then released to the public. Later, when an attacker collects and uses y_{prt} to generate customized images by M, the original ID ori cannot be identified from those customized images. In summary, the user's objective is to create a protected image y_{prt} that can degrade the efficacy of ID customization.

Specifically, the attacker can use different tuning-free models for ID customization. Therefore, the image protector needs to craft a perturbation δ that can achieve universal protection on various tuning-free models. Additionally, the protection should be robust to image transformations that can weaken the protection efficacy.

IV. Method

Fig. 1 shows the framework of FaceOff. To suppress the original ID information in customized images, we employ contrastive learning to draw the protected image closer to the target image while pushing it away from the original image in the semantic space. To mitigate performance degradation caused by image transformations, we further integrate Gaussian filtering into the optimization process.

A. Contrastive Loss

To achieve universal protection, we disturb the image encoders of three tuning-free models (i.e., IP-Adapter, IP-Adapter Plus, and PhotoMaker), deviating the semantics of protected images away from those of the original images:

$$L_{dev} = \sum_{i=1}^{N} cos(\mathcal{E}_i(y_{prt}), \mathcal{E}_i(y_{ori})), \tag{1}$$

TABLE I
COMPARISON OF BASELINE AND FACEOFF ACROSS DIFFERENT MODELS ON VGGFACE2 AND CELEBA-HQ.

Dataset	Method	IP-Adapter		IP-Adapter Plus		PhotoMaker		Average		Imperceptibility	
		FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓	LPIPS↓	PSNR↑
VGGFace2	No Defense	0.034	0.328	0.000	0.433	0.138	0.436	0.057	0.399	-	-
	IDProtector [20]	0.260	0.128	0.001	0.159	0.287	0.110	0.183	0.132	0.295	27.304
	FaceOff	**0.546**	**0.084**	**0.046**	**0.086**	**0.344**	**0.048**	**0.312**	**0.073**	**0.241**	**27.935**
CelebA-HQ	No Defense	0.051	0.322	0.000	0.466	0.128	0.455	0.060	0.414	-	-
	IDProtector [20]	0.394	0.102	0.000	0.106	0.262	0.050	0.219	0.086	0.298	27.232
	FaceOff	**0.646**	**0.056**	**0.030**	**0.069**	**0.354**	**0.029**	**0.343**	**0.051**	**0.237**	**27.945**

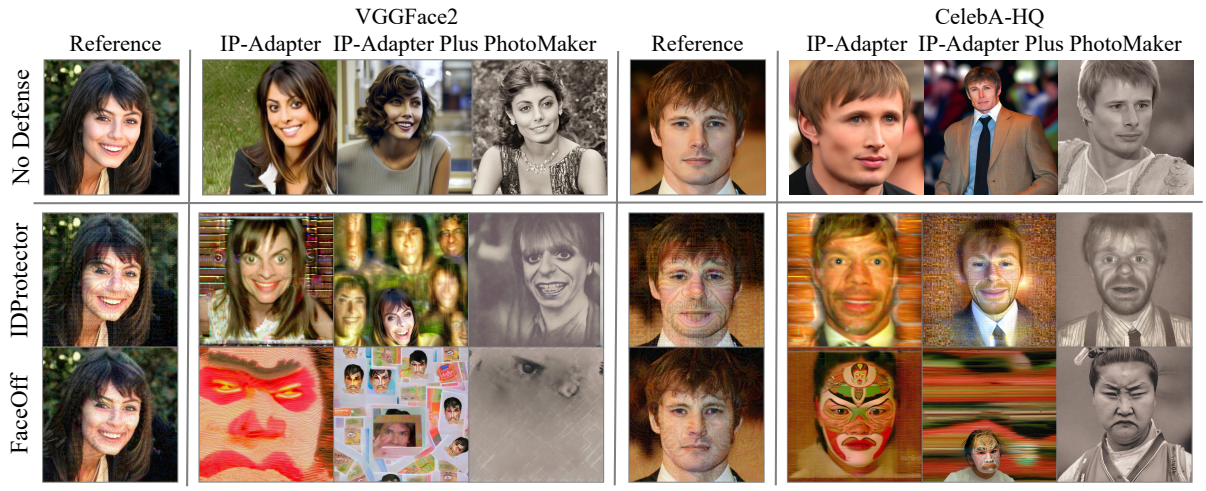

Fig. 2. Qualitative comparison with baseline. The inference prompt is "a photo of _ person".

where $cos(\cdot)$ is the cosine similarity, $\mathcal{E}_i$ is the image encoder of i-th customization model, N is the number of customization models, y_{prt} is the protecting image, and y_{ori} is the original image. However, customized images after protection remain recognized human faces, which may infringe on the copyrighted portrait images. Therefore, we design a contrastive loss to reduce the identified human faces in the customized images. Specifically, we shift the semantics of the protected image closer to those of the target image while pushing them away from the semantics of the original image. The contrastive loss is formulated as follows:

$$L_{con} = \sum_{i=1}^{N} cos(\mathcal{E}_i(y_{prt}), \mathcal{E}_i(y_{ori})) - cos(\mathcal{E}_i(y_{prt}), \mathcal{E}_i(y_{tgt})), \tag{2}$$

where y_{tgt} is the target image. By optimizing the contrastive loss, we obtain protective perturbations that reduce the identified human faces in customized images and mitigate copyright infringement risk.

B. Gaussian Augmentation

However, the effectiveness of such perturbations may degrade under image transformations. In particular, operations such as JPEG compression suppress high-frequency noise that deviates from natural image statistics, thereby significantly weakening adversarial perturbations that primarily rely on high-frequency components. To improve robustness against image transformations, we incorporate Gaussian filtering into the perturbation optimization process. The Gaussian augmentation loss is formulated as:

$$\begin{aligned} L_{gau} = \sum_{i=1}^{N} &\cos(\mathcal{E}_i(G(y_{prt})), \mathcal{E}_i(G(y_{ori}))) \\ &- \cos(\mathcal{E}_i(G(y_{prt})), \mathcal{E}_i(G(y_{tgt}))), \end{aligned} \tag{3}$$

where $G(\cdot)$ is the Gaussian filtering transformation. This design explicitly simulates the suppression of high-frequency components, encouraging the optimization to produce perturbations dominated by low- and mid-frequency information that remains effective after transformation.

Although the Gaussian augmentation loss improves protection efficacy under image transformations, it reduces efficacy under standard customization. To this end, we combine the contrastive loss with the Gaussian augmentation loss to preserve protection efficacy under standard customization, yielding the overall FaceOff loss:

$$L_{FaceOff} = L_{con} + L_{gau}. \tag{4}$$

Specially, we utilize the PGD algorithm [32] to optimize the perturbation by minimizing FaceOff loss:

$$\delta := \arg\min_{\delta} L_{FaceOff}, s.t.\ ||\delta|| \leq \eta, \tag{5}$$

where δ is the perturbation and η is the noise budget.

TABLE II
ROBUSTNESS OF FACEOFF UNDER IMAGE TRANSFORMATIONS ON VGGFACE2.

Setting	IP-Adapter		IP-Adapter Plus		PhotoMaker	
	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓
No Defense	0.034	0.328	0.000	0.433	0.138	0.436
SC	0.546	0.084	0.046	0.086	0.344	0.048
JPEG [33]	0.448	0.091	0.002	0.146	0.277	0.114
GrIDPure [21]	0.233	0.098	0.008	0.144	0.283	0.107
RM [22]	0.097	0.220	0.006	0.281	0.220	0.253

TABLE III
ABLATION STUDY OF FACEOFF ON VGGFACE2 UNDER JPEG.

Con.	Aug.	IP-Adapter		IP-Adapter Plus		PhotoMaker		Average	
		FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓
×	×	0.185	0.116	0.001	0.162	0.202	0.150	0.129	0.143
✓	×	0.298	0.116	**0.002**	0.183	0.226	0.188	0.175	0.162
✓	✓	**0.448**	**0.091**	**0.002**	**0.146**	**0.277**	**0.114**	**0.242**	**0.117**

V. EXPERIMENTS

A. Experimental Setup

Datasets. Following previous work [13], [20], we evaluate FaceOff on two facial benchmarks: VGGFace2 [34] and CelebA-HQ [35]. We select 400 images with 50 IDs for each dataset. All images are resized to a resolution of 224 × 224.

Target Models. We test FaceOff on three tuning-free models: IP-Adapter [8], IP-Adapter Plus [8], and PhotoMaker [10]. The base diffusion model is SDXL base 1.0 [3].

Metrics. Following previous works [13], [20], we adopt two metrics: Face Detection Failure Rate (FDFR) and Identity Score Matching (ISM). We first use a RetinaFace detector to detect faces in customized images, and calculate the rate of customized images that have no detectable face, denoted as FDFR. Then, if a face is detected, we use the ArcFace recognizer [36] to extract its face recognition embedding and compute its cosine distance to the average face embedding of the original images, denoted as ISM.

Implementation Details. We set PGD iteration rounds to 100. We set the step size α as 0.005 and the noise budget η as $16/255$ following [37]. The target image is the Beijing Opera Facial Mask of the role Yingbu. During inference, we leverage two prompts "a photo of sks person" and "a DSLR portrait of sks person" in [13] for PhotoMaker. The "sks" is replaced with an empty string for IP-Adapter and IP-Adapter Plus. Each prompt is used to generate 16 images.

B. Comparison with Baseline

We compare FaceOff with a SOTA baseline IDProtector [20]. In Tab. I, IDProtector achieves a significant reduction in ISM, but its performance on FDFR still requires further improvement. FaceOff obtains the highest FDFR and the lowest ISM values on average, indicating that the customized image protected by FaceOff has the fewest detectable faces and the lowest similarity between the detected faces and the original faces. Moreover, Tab. I shows that FaceOff achieves better perturbation imperceptibility than IDProtector, reflecting superior visual quathe lity of the protected images. Fig. 2 shows that FaceOff effectively distorts the original ID in customized images, reducing the proportion of recognizable faces in customized images. In contrast, IDProtector produces customized images with many recognizable human faces, potentially infringing on copyrighted portrait images.

TABLE IV
TRANSFERABILITY OF FACEOFF ON UNSEEN MODELS.

Method	IP-Adapter		IP-Adapter Plus		Fastcomposer		Face-diffuser	
	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓
No Defense	0.033	0.340	0.031	0.369	0.015	0.383	0.000	0.386
FaceOff	0.353	0.078	0.309	0.081	0.001	0.080	0.004	0.097

C. Robustness against image transformations

In Tab. II, we evaluate the robustness of FaceOff under three image transformations, including JPEG compression [33] and two adversarial purification methods, GrIDPure [21] and Robust-Mimicry (RM) [22]. The standard customization pipeline without any additional image transformation is denoted as SC. In Tab. II, under all image transformations, the customization performance can not be restored to the No Defense setting. FaceOff exhibits strong resistance to JPEG and GrIDPure, maintaining a high FDFR and a low ISM. RM leads to a more pronounced protection degradation of FaceOff, which can be attributed to the image super-resolution diffusion model employed by RM that effectively removes adversarial perturbations. Even so, the ISM under RM remains substantially lower than that of the No Defense setting.

D. Transferability on unseen models

In Tab. IV, we evaluate the transferability of FaceOff on unseen base model (SD v1.5) and customization methods (FastComposer and Face-Diffuser). Specifically, for IP-Adapter and IP-Adapter Plus, their base model is replaced from the SDXL base 1.0 to SD v1.5. FastComposer and Face-Diffuser are two customization methods built upon SD v1.5. In Tab. IV, images protected by FaceOff consistently lead to a substantial increase in FDFR and a significant reduction in ISM. On average, the FDFR increases from 0.020 to 0.167, while the ISM decreases from 0.370 to 0.084. These results demonstrate strong transferability of FaceOff on unseen models. FaceOff requires about 19 GB of VRAM and 240 s to optimize a single ID. Unless otherwise specified,weconduct experimentsunder standard customizationwithout imagetransformations.

E. Ablation Study

In Tab. III, we conduct ablation studies on the VGGFace2 dataset under JPEG compression to assess the contribution of each FaceOff module. The contrastive loss module raises the average FDFR from 12.9% to 17.5%, meaning fewer faces can be recognized in customized images. The Gaussian augmentation module further improves the average FDFR from 17.5% to

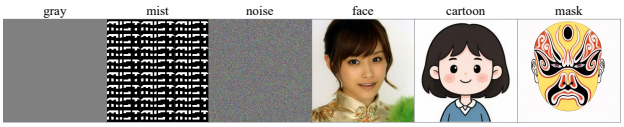

Fig. 3. Different target images.

TABLE V
EFFECT OF DIFFERENT TARGET IMAGES FOR FACEOFF ON VGGFACE2. BOLD: BEST; UNDERLINE: SECOND BEST.

Setting	IP-Adapter		IP-Adapter Plus		PhotoMaker		Average	
	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓
No Defense	0.034	0.328	0.000	0.433	0.138	0.436	0.057	0.399
gray	0.156	0.177	0.000	0.149	0.249	0.177	0.135	0.168
mist	0.511	0.119	0.054	0.163	0.281	0.163	0.282	0.148
noise	**0.673**	0.130	0.043	0.182	0.135	0.204	0.284	0.172
face	0.091	0.088	0.002	0.121	0.240	**0.078**	0.111	**0.096**
cartoon	0.445	0.095	**0.141**	0.177	0.157	0.180	0.248	0.151
mask	0.544	**0.077**	0.106	**0.115**	**0.419**	0.100	**0.356**	0.097

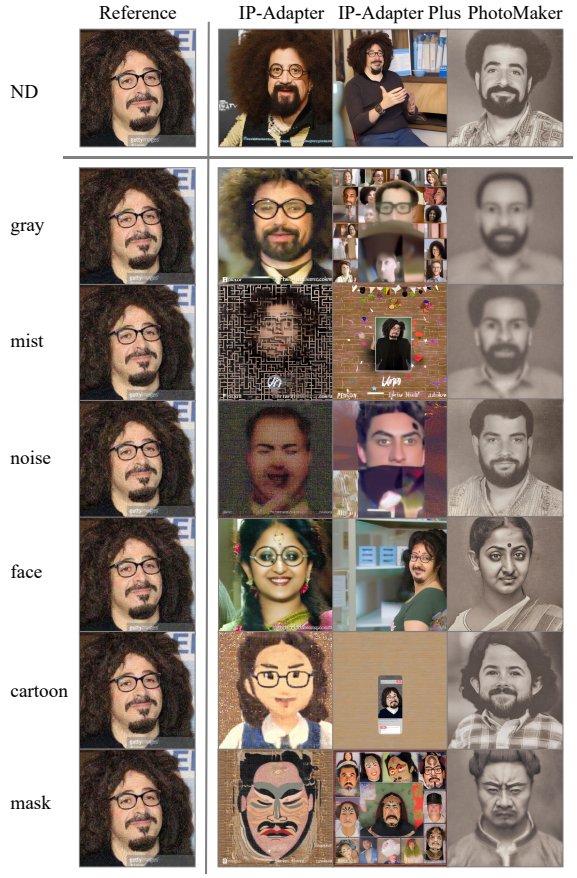

Fig. 4. Visualization of FaceOff with different target images.

24.2% while reducing the average ISM from 16.2% to 11.7%, demonstrating stronger protection under JPEG compression. Combining all modules, FaceOff achieves the highest FDFR and the lowest ISM, confirming its robust protection.

F. Parameter Analysis

Target Images. In Tab. V, we evaluate the impact of different target images. The evaluated target images are shown in Fig. 3. Specifically, $gray$ and $mist$ are used in existing methods for artistic copyright protection [38], [39]; $noise$ denotes a random noise image; $face$ is chosen by computing the cosine similarity of CLIP features among images of different IDs in the VGGFace2 dataset, and for each ID, selecting an image of another ID with the lowest similarity as the target; $cartoon$ represents a cartoon-style image; $mask$ corresponds to the facial mask of the Peking Opera character Yingbu. $mask$ achieves the best FDFR and the second-best ISM. This can be attributed to the fact that tuning-free models are adapted to generate portrait images. $mask$ preserves facial structure while substantially deviating from real-face semantics, thereby more effectively disrupting ID customization. We observe that $face$ achieves the lowest average ISM, indicating the strongest protection. However, this setting requires selecting a distinct target face for each ID and ensuring the absence of copyright issues, leading to substantial overhead. Consequently, we adopt $mask$ as the default target image instead of $face$. Fig. 4 visualizes customized images generated under different target images. The protected customized images show substantial ID distortion and align with the target characteristics. Notably, using $mask$ or $face$ yields the lowest ID similarity.

Augmentation Strategies. In Tab. VI, we analyze the effects of different image augmentation strategies, including JPEG compression, ColorJitter, Gaussian noise, and Gaussian filtering. All strategies are evaluated under two settings: standard customization and JPEG compression. The results show that all augmentation strategies effectively increase the FDFR and reduce the ISM, demonstrating the flexibility of FaceOff. Specifically, Gaussian filtering achieves the best balance under both standard customization and JPEG compression. By suppressing high-frequency components while preserving low- and mid-frequency perturbations, it enhances perturbation stability. Its low-pass nature also aligns with JPEG, which retains low-frequency content and removes high-frequency details, making the perturbations more robust to compression.

VI. CONCLUSION

This paper proposes FaceOff, a framework that protects user portrait images from tuning-free ID customization. We address the limitations of prior works in copyright infringement risk and fragility to image transformations through contrastive loss and Gaussian augmentation. Extensive experiments demonstrate that FaceOff can effectively reduce identified ID in customized images and enhance protection robustness against image transformations.

REFERENCES

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

TABLE VI
EFFECT OF DIFFERENT AUGMENTATION STRATEGIES FOR FACEOFF ON VGGFACE2. BOLD: BEST; UNDERLINE: SECOND BEST.

Setting	Augmentation Strategy	IP-Adapter		IP-Adapter Plus		PhotoMaker		Average	
		FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓	FDFR↑	ISM↓
Standard Customization	JPEG	0.612	0.085	0.093	**0.084**	0.208	0.049	0.304	0.073
	ColorJitter	0.646	0.085	**0.118**	**0.084**	0.214	**0.045**	**0.326**	**0.071**
	Gaussian Noise	**0.673**	**0.084**	0.062	0.100	0.191	0.057	0.309	0.080
	Gaussian Filtering	0.546	**0.084**	0.046	0.086	**0.344**	0.048	0.312	0.073
JPEG	JPEG	0.419	0.118	0.001	0.201	0.151	0.246	0.190	0.188
	ColorJitter	0.409	0.116	0.001	0.197	0.153	0.253	0.188	0.189
	Gaussian Noise	0.374	0.129	0.001	0.211	0.156	0.271	0.177	0.204
	Gaussian Filtering	**0.448**	**0.091**	**0.002**	**0.146**	**0.277**	**0.114**	**0.242**	**0.117**

[3] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *ICLR*, 2024.

[4] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[5] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023, pp. 22 500–22 510.

[6] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *CVPR*, 2023, pp. 1931–1941.

[7] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman, "Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models," in *CVPR*, 2024, pp. 6527–6536.

[8] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[9] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han, "Fastcomposer: Tuning-free multi-subject image generation with localized attention," *arXiv preprint arXiv:2305.10431*, 2023.

[10] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, "Photomaker: Customizing realistic human photos via stacked id embedding," in *CVPR*, 2024, pp. 8640–8650.

[11] W. Chen, H. Hu, Y. Li, N. Ruiz, X. Jia, M.-W. Chang, and W. W. Cohen, "Subject-driven text-to-image generation via apprenticeship learning," *NeurIPS*, vol. 36, 2024.

[12] Z. Guo, Y. Wu, C. Zhuowei, P. Zhang, Q. He *et al.*, "Pulid: Pure and lightning id customization via contrastive alignment," *NeurIPS*, vol. 37, pp. 36 777–36 804, 2024.

[13] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, and A. Tran, "Anti-dreambooth: Protecting users from personalized text-to-image synthesis," in *ICCV*, 2023, pp. 2116–2127.

[14] Y. Liu, C. Fan, Y. Dai, X. Chen, P. Zhou, and L. Sun, "Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning," in *CVPR*, 2024, pp. 24 219–24 228.

[15] F. Wang, Z. Tan, T. Wei, Y. Wu, and Q. Huang, "Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models," in *CVPR*, 2024, pp. 12 047–12 056.

[16] J. Xu, Y. Lu, Y. Li, S. Lu, D. Wang, and X. Wei, "Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models," in *CVPR*, 2024, pp. 24 534–24 543.

[17] Y. Liu, J. An, W. Zhang, D. Wu, J. Gu, Z. Lin, and W. Wang, "Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization," in *MM*, 2024, p. 3587–3596.

[18] C. Wan, Y. He, X. Song, and Y. Gong, "Prompt-agnostic adversarial perturbation for customized diffusion models," *NeurIPS*, vol. 37, pp. 136 576–136 619, 2024.

[19] B. Zheng, C. Liang, and X. Wu, "Targeted attack improves protection against unauthorized diffusion customization," in *ICLR*, 2025.

[20] Y. Song, P. Yang, H. Ci, and M. Z. Shou, "Idprotector: An adversarial noise encoder to protect against id-preserving image generation," *CVPR*, 2025.

[21] Z. Zhao, J. Duan, K. Xu, C. Wang, R. Zhang, Z. Du, Q. Guo, and X. Hu, "Can protective perturbation safeguard personal data from being exploited by stable diffusion?" in *CVPR*, 2024, pp. 24 398–24 407.

[22] R. Hönig, J. Rando, N. Carlini, and F. Tramèr, "Adversarial perturbations cannot reliably protect artists from generative ai," *ICLR*, 2025.

[23] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang, "Svdiff: Compact parameter space for diffusion fine-tuning," in *ICCV*, 2023, pp. 7323–7334.

[24] G. Yuan, X. Cun, Y. Zhang, M. Li, C. Qi, X. Wang, Y. Shan, and H. Zheng, "Inserting anybody in diffusion models via celeb basis," *arXiv preprint arXiv:2306.00926*, 2023.

[25] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo, "Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation," in *ICCV*, 2023, pp. 15 943–15 953.

[26] J. Ma, J. Liang, C. Chen, and H. Lu, "Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.

[27] J. Shi, W. Xiong, Z. Lin, and H. J. Jung, "Instantbooth: Personalized text-to-image generation without test-time finetuning," in *CVPR*, 2024, pp. 8543–8552.

[28] X. Peng, J. Zhu, B. Jiang, Y. Tai, D. Luo, J. Zhang, W. Lin, T. Jin, C. Wang, and R. Ji, "Portraitbooth: A versatile portrait model for fast identity-preserved personalization," in *CVPR*, 2024, pp. 27 070–27 080.

[29] Y. Wang, W. Zhang, J. Zheng, and C. Jin, "High-fidelity person-centric subject-to-image synthesis," in *CVPR*, 2024, pp. 7675–7684.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[31] N. Ahn, K. Yoo, W. Ahn, D. Kim, and S.-H. Nam, "Nearly zero-cost protection against mimicry by personalized diffusion models," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

[32] A. Madry, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[33] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: Dnn-oriented jpeg compression against adversarial examples," in *CVPR*. IEEE, 2019, pp. 860–868.

[34] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[35] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015, pp. 3730–3738.

[36] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[37] H. Xue, C. Liang, X. Wu, and Y. Chen, "Toward effective protection against diffusion-based mimicry through score distillation," in *ICLR*, 2024.

[38] H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Madry, "Raising the cost of malicious ai-powered image editing," in *ICML*, 2023.

[39] C. Liang and X. Wu, "Mist: Towards improved adversarial examples for diffusion models," *arXiv preprint arXiv:2305.12683*, 2023.