Countering Personalized Text-to-Image Generation with Influence Watermarks

Supplementary Material

A. Proofs

Proposition 1. *To confine the optimization to the low frequency subspace, directly masking the gradient is equivalent to mask the image and perform gradient-based optimization through discrete cosine transform.*

Proof. For the reference image x_0 and its corresponding frequency matrix, we have $X_0(k,\cdot,\cdot) = \mathrm{DCT}(x_0(k,\cdot,\cdot))$ and $x_0(k,\cdot,\cdot) = \mathrm{IDCT}(X_0(k,\cdot,\cdot))$. Considering the vector $\mathbf{z}$, its right-product with the Jacobian of IDCT can be calculated by $J_{\mathrm{IDCT}} \cdot \mathbf{z} = \mathrm{DCT}(\mathbf{z})$. According to the chain rule, we have:

$$\frac{\partial \mathrm{SE}_{\epsilon,\theta^*,t,c}(x)}{\partial X_0} = \mathrm{DCT}(\frac{\partial \mathrm{SE}_{\epsilon,\theta^*,t,c}(x)}{\partial x_0}). \tag{1}$$

Therefore, directly masking the gradient is equivalent to mask the image and perform gradient-based optimization through discrete cosine transform. □

Proposition 2. *To forge unlearnable example $\tilde{x}_0$ from x_0 against personalized diffusion models, the self-influence $\mathcal{I}(x_0)$ can be inferred as:*

$$\mathcal{I}(x_0) = -\nabla_\theta \mathrm{SE}_{\epsilon,\theta^*,t,c}(x_0)^{\mathrm{T}} H_{\theta^*}^{-1} \nabla_{x_0} \nabla_\theta \mathrm{SE}_{\epsilon,\theta^*,t,c}(x_0), \tag{2}$$

where $H_{\theta^*} = \mathbb{E}_{x_0}[\nabla_\theta^2 \mathrm{SE}_{\epsilon,\theta^*,t,c}(x_0)]$ is the Hessian and we assume that H_{θ^*} is positive definite.

Proof. For simplicity, we denote $L(x,\theta) = \mathrm{SE}_{\epsilon,\theta,t,c}(x)$, and θ^* and θ_δ^* as the empirical risk minimizer and empirical risk minimizer for the perturbed image, respectively:

$$\theta^* = \arg\min_\theta L(x_0,\theta),$$
$$\theta_\delta^* = \arg\min_\theta L(x_0+\delta,\theta).$$

Then we can rewrite $\mathcal{I}(x_0) = \nabla_\delta \mathrm{SE}_{\epsilon,\theta_\delta^*,t,c}(x_0+\delta)\big|_{\delta=0} = \nabla_\delta L(x_0+\delta,\theta_\delta^*)\Big|_{\delta=0}$. And next we have to prove that $\mathcal{I}(x_0) = -\nabla_\theta L(x_0,\theta^*)^{\mathrm{T}} H_{\theta^*}^{-1} \nabla_{x_0} \nabla_\theta L(x_0,\theta^*)$.

In accordance with Koh *et al.* [3], the perturbed minimizer θ_δ^* can be approximated by moving a unit mass from x to $x+\delta$ *w.r.t.* θ^*. Utilizing the upweighting influence in Lemma 1, this leads to the following expression:

$$\begin{aligned} \theta_\delta^* - \theta^* &\approx \mathcal{I}'(x_0+\delta) - \mathcal{I}'(x_0) \\ &= -H_{\theta^*}^{-1}(\nabla_\theta L(x_0+\delta,\theta^*) - \nabla_\theta L(x_0,\theta^*)) \\ &\approx -H_{\theta^*}^{-1} \nabla_{x_0} \nabla_\theta L(x_0,\theta^*)\delta, \end{aligned} \tag{3}$$

where in the last step, we assume that δ is sufficiently small (for the image quality of $x_0+\delta$). It then derives:

$$\left.\frac{\mathrm{d}\theta_\delta^*}{\mathrm{d}\delta}\right|_{\delta=0} = -H_{\theta^*}^{-1} \nabla_{x_0} \nabla_\theta L(x_0,\theta^*). \tag{4}$$

Finally, we apply the chain rule and substitute this result:

$$\begin{aligned} \mathcal{I}(x_0) &= \nabla_\delta L(x_0+\delta,\theta_\delta^*)\Big|_{\delta=0} \\ &= \nabla_\theta L(x_0,\theta^*)^{\mathrm{T}} \left.\frac{\mathrm{d}\theta_\delta^*}{\mathrm{d}\delta}\right|_{\delta=0} \\ &= -\nabla_\theta L(x_0,\theta^*)^{\mathrm{T}} H_{\theta^*}^{-1} \nabla_{x_0} \nabla_\theta L(x_0,\theta^*). \end{aligned} \tag{5}$$

□

Table 1. Possible modifications on watermarked images. $^\uparrow$ indicates that the higher values represent better performance.

	BRISQUE (↑)		
	photo	*portrait*	**Avg.**
Clean	18.61	2.10	10.36
JPEG$_{q=0.95}$	35.85	54.60	45.23
JPEG$_{q=0.75}$	49.53	36.54	43.04
WeChat [4]	33.66	21.72	27.69

Lemma 1 (Upweighting influence [2, 3])**.** *Given the empirical risk minimizer θ_{ϵ,x_0}^* with ϵ upweighted perturbations. The influence from upweighting x_0 on parameters θ^* yields:*

$$\mathcal{I}'(x_0) = \left.\frac{\mathrm{d}\theta_{\epsilon,x_0}^*}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = -H_{\theta^*}^{-1} \nabla_\theta L(x_0,\theta^*), \tag{6}$$

where $H_{\theta^*} = \mathbb{E}_{x_0}[\nabla_\theta^2 \mathrm{SE}_{\epsilon,\theta^*,t,c}(x_0)]$ is the Hessian assumed to be positive definite.

B. Possible Modifications

We provide additional experiments on reference images under possible modifications. In this part, different JPEG quality factors are considered, and we also validate the robustness of our proposed method on real social platforms. We report empirical results in Tab. 1. Under all possible modifications, the generated images suffer obvious deficits in visual quality, when the reference images are protected by our proposed method. It is observed that our proposed method is still robust against various compression quality factors, and on real social platforms like WeChat [4].

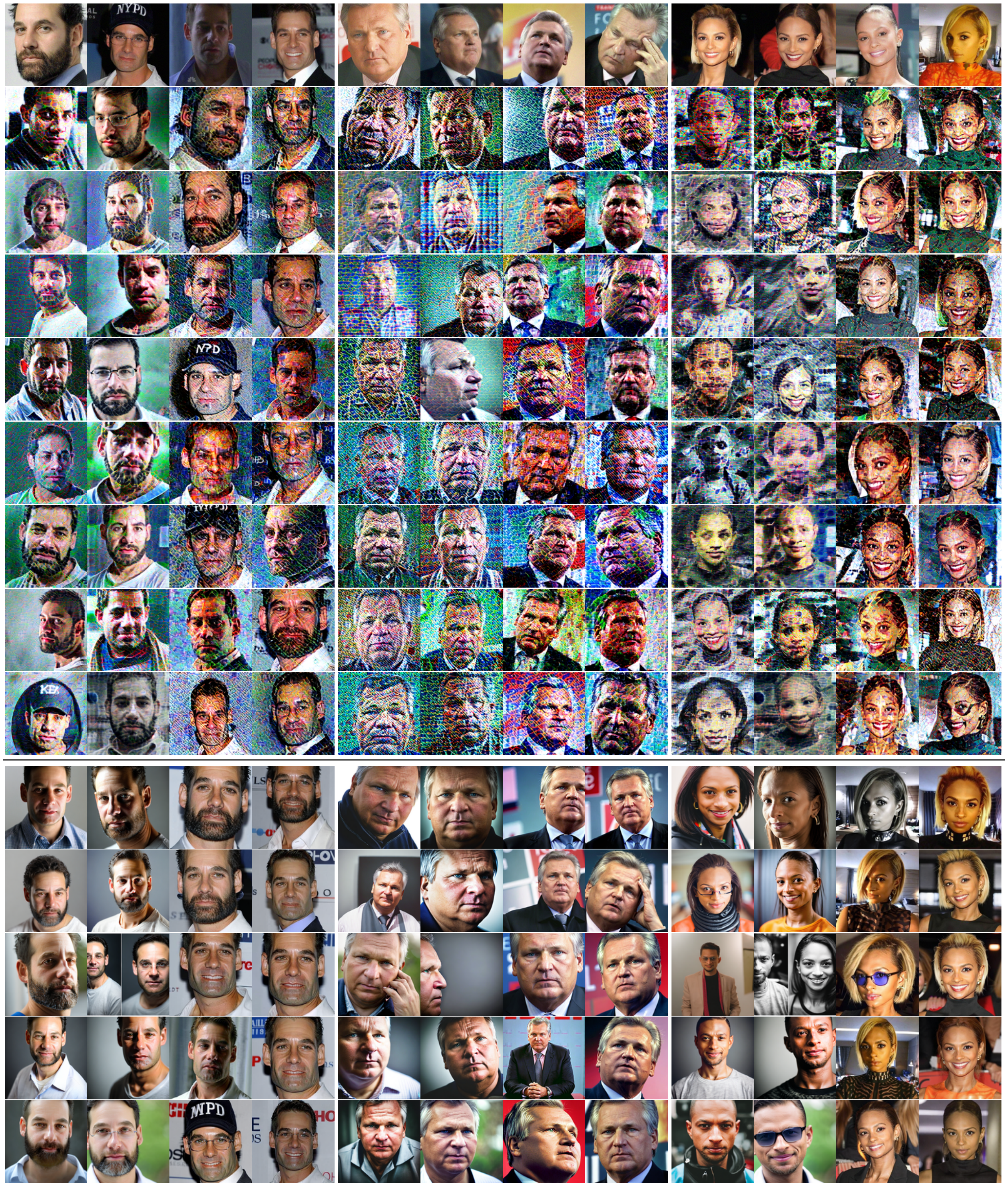

Figure 1. Additional visual results on VGGFace2. Reference images with `InMark` are present at the *first row*. The generated images in the *last 5 rows* represent personalized results from clean reference images.

C. Additional Visual Results

In this part, we present more visual results in Fig. 1 on VGGFace2 [1] when reference images are protected by our proposed InMark, where DreamBooth [5] is used for personalized text-to-image generation. The results show that most of our generated images suffer from a significant loss of texture and pattern quality compared to the reference images. Even in the very few cases where the facial appearance is preserved, the overall colors are distorted, making them useless for fake news propagation. Therefore, it is empirically proven that our proposed InMark can destroy the generated image quality in detail.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pages 67–74, 2018. 3

[2] R Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980. 1

[3] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, pages 1885–1894, 2017. 1

[4] Jiezhong Qiu, Yixuan Li, Jie Tang, Zheng Lu, Hao Ye, Bo Chen, Qiang Yang, and John E. Hopcroft. The lifecycle and cascade of wechat social messaging groups. In *WWW*, pages 311–320. ACM, 2016. 1

[5] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 3