My Art My Choice: Adversarial Protection Against Unruly AI

Anthony Rhodes
Intel Labs

Ram Bhagat
Binghamton University

Umur Aybars Çiftçi
Binghamton University

İlke Demir
Intel Labs

{anthony.rhodes,ilke.demir}@intel.com, {rbhagat2,uciftci}@binghamton.edu

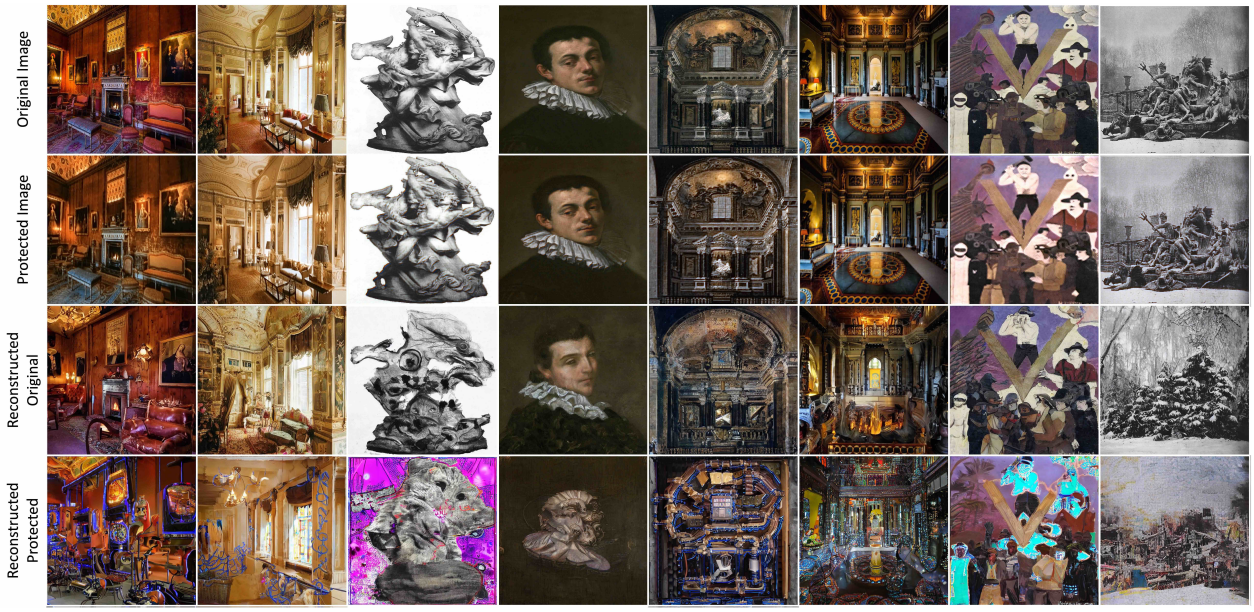

Figure 1. Our approach enables artists to protect their content (first row) by learning to create perturbed versions (second row). Diffusion models exploit the original artwork (third row), however, protected images break these models (last row).

Abstract

Generative AI is on the rise, enabling everyone to produce realistic content via publicly available interfaces. Especially for guided image generation, diffusion models are changing the creator economy by producing high-quality low-cost content. In parallel, artists are rising against unruly AI, since their artwork is leveraged, distributed, and dissimulated by large generative models. My Art My Choice (MAMC) aims to empower content owners by protecting their copyrighted materials from being utilized by diffusion models in an adversarial fashion. MAMC learns to generate adversarially perturbed "protected" versions of images which can in turn "break" diffusion models. The perturbation amount is decided by the artist to balance distortion vs. protection of the content. We experiment on four datasets, both protected image and diffusion output results are evaluated in visual, noise, structure, pixel, and generative spaces.

1. Introduction

Generative modeling has been introduced over half a century ago, with applications in mathematics [30], shapes [8], botany [1], architecture [7], and many other domains. Recently, deep counterparts of generative models are proliferating as a realistic way of creating visual content. They are able to mimic specific content, style, or structure of training samples – replicating art. Consequently, artists and creators are resisting against unruly use of AI [3, 9, 14], since (1) generative AI can create derivatives of their art without any liabilities, (2) diffusion models are trained on their data without their permission, and (3) there is no compensation mechanism for their replicated art.

Because regulation and policy are insufficiently mature to provide robust means to protect creative rights, we propose an interim AI tool to enable artists to seal their material with adversarial protection against generative AI systems. "My Art My Choice" provides artists with the ability to protect their content from being used in generative AI

applications, in addition to empowering users to decide on the specific strength of protection suitable for their use case. Our contributions include:

- A model that learns to adversarially protect any given image against diffusion models,
- A human-centric AI system to balance protection and perturbation of the content, and
- Experimental validation of the approach in various generative AI tasks and datasets.

Our results span four datasets, qualitatively and quantitatively comparing (1) input and protected images, and (2) diffusion model outputs of input and protected images, as demonstrated in Fig.1. We also compare to image cloaking techniques and explore different application domains.

2. Related Work

2.1. Controlled Content Generation

DMs learn generating specific identity, style, or context with a few input images via fine-tuning [15, 20] or by tokenization [29], as image guidance emerges for DMs. Recently, DMs are used for crafting stories [13], deepfakes [21], multi-person images [37], pasting objects into scenes [38, 41], image editing [4, 16, 39, 43], object editing [11], video synthesis [2], 3D avatars [36], and many other applications; all of which require additional image input for guidance. MAMC aims to break these models by protecting this guiding image.

2.2. Adversarial Generation

There has been a considerable amount on literature for using image manipulation and generation in adversarial settings [32], including those against face recognition [5, 6, 26], breaking deepfake detectors [24], and preemptively confusing models by data augmentation [10]. These approaches utilize generative models to attack other deep learning models, however, in adversarial protection, the loss is guided by the black box model's response to adversarially perturbed inputs. To clarify this distinction, MAMC attacks synthesis models, whereas others attack analysis models. This adds another layer of complexity to define our threat model, i.e., what it means to 'break a diffusion model' vs. 'break a face recognizer', as it will be explained later.

2.3. Adversarial Protection

Most of the aforementioned DMs are trained on large internet scraped visual datasets such as LiASON [25] without any ownership or copyright monitoring. As a result, DMs can replicate the content [31], style [42], and structure [39] of samples; which is violating artists' rights over their own materials. Emerging research addresses this problem by machine unlearning [33], by confusing the model to converge towards a different style target [28], by focusing on disabling specific DMs [35] such as [20], by injecting noise into input images [23] to disable text-based editing, and by Compartmentalized Diffusion Models [12] to selectively forget by continual learning. MAMC follows this route to provide adversarial protection: (1) by learning to generate imperceptible adversarial twins, (2) using a combination of losses robust against several tasks, and (3) with external controllable balancing between distortion and protection of an image. Unlike previous work like [28], (1) MAMC does not need a driving image, (2) MAMC lets the artist set the amount of distortion, (3) MAMC defines a multi-objective training regime for higher quality and more atrophy, and (4) MAMC is not limited to specific tasks, models, or domains.

3. My Art My Choice

Given an image I, MAMC learns to generate $G(I) = I + \delta = I'$ where δ is the learned perturbation to attack a black-box diffusion model M. This attack should create an image as dissimilar to the expectations as possible, meaning that $M(I)$ and $M(I')$ should be maximally dissimilar. Thus, adversarial protection optimization becomes,

$$\max_{\delta_I} ||M(I+\delta_I) - M(I)||, \;\; s.t. |\delta_I| < \phi_I + \epsilon \quad (1)$$

where ϕ is the balance factor and ϵ is a small neighborhood.

3.1. Architecture

We employ a simple UNet architecture [18] to learn this generation process, consisting of blocks with two convolutional layers followed by up/downsampling, with concatenations between every encoder/decoder block (Fig. 2). We use a standard pre-trained diffusion model [17] in frozen state to infer input output relations.

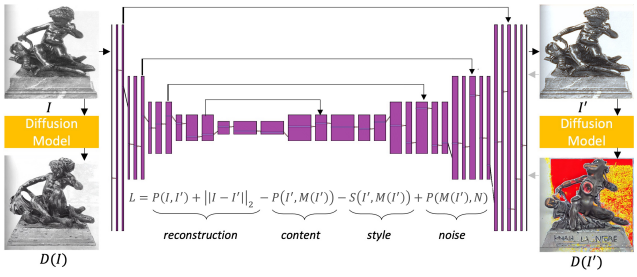

Figure 2. Our input and output samples, generator architecture, and loss formulation is simplified in this overview.

3.2. Training Objective

Artists expect minimal changes in their artwork, thus we introduce a reconstruction term L_R. We use LPIPS [40] for perceptual similarity $\mathcal{P}$. We also add a pixel-wise ℓ_2 norm to prevent color shifts.

$$L_R = \alpha_{R1}\mathcal{P}(I, I') + \alpha_{R2}||I - I'||_2^2 \quad (2)$$

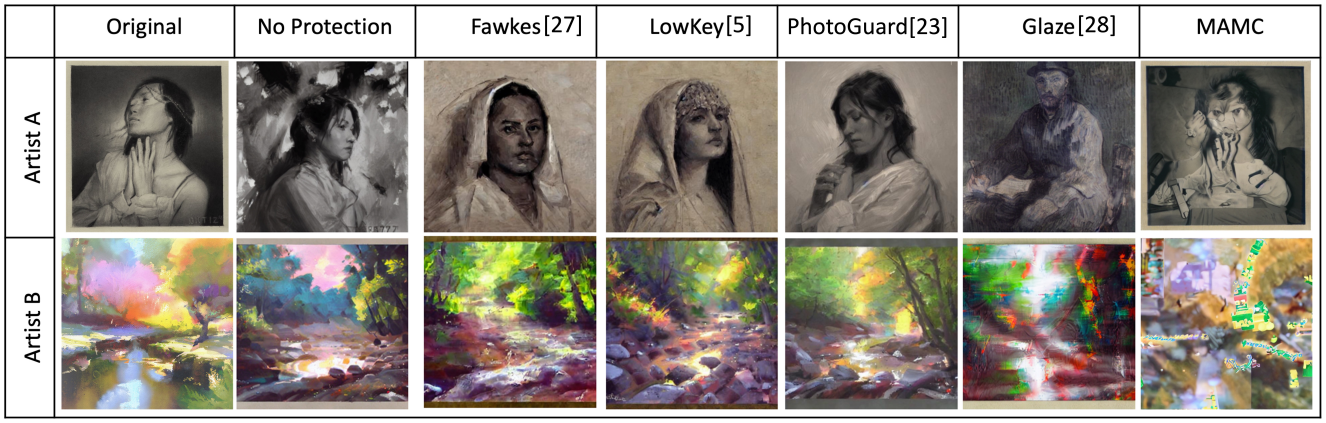

Figure 3. MAMC visually compared to cloaking approaches, output is so distorted that it is obviously not useful, in the style of nobody.

Especially to protect against inpainting and personalization, we introduce a content loss where the diffusion output is perceptually dissimilar to the protected image.

$$L_C = -\alpha_C \mathcal{P}(I', M(I')) \tag{3}$$

To prohibit style transfer and reconstruction, we introduce a style loss as the distance between Gram Matrices Ω of the protected image and its diffusion output, over activations j.

$$L_S = -\alpha_S \frac{1}{|j|} \sum_j ||\Omega_j(I') - \Omega_j(M(I'))|| \tag{4}$$

Finally, to confuse the diffusion model, we introduce a noise loss to put diffusion output of the protected image towards Gaussian noise, indicated by $\mathcal{N}$.

$$L_N = \alpha_{N1} \mathcal{P}(M(I'), \mathcal{N}) \tag{5}$$

Our overall loss function is constructed as follows setting weights α_* experimentally.

$$L = \alpha_R L_R - \alpha_C L_C - \alpha_S L_S + \alpha_N L_N \tag{6}$$

3.3. Balance Factor

We would like to provide control to the users, especially as MAMC alters their materials. We experiment with predefined α_* values as MAMC with different strengths. These models are then exposed to balance distortion (higher α_R) vs. protection (higher α_N) of their images (Sec. 4.4).

4. Results

We present evaluations and experiments of My Art My Choice on four datasets for a comprehensive understanding across domains: Wiki Art [22] with 1K and 5K subsets, Historic Art [34] with 1K and 5K subsets, single artist datasets with 200 images, and FaceForensics++ [19] with 100 images. We select these as representative datasets, covering diverse content, style, artist, and domain fronts.

We want to validate that (1) input and protected images are similar enough and (2) diffusion output of the protected image has low quality. We visualize the success of MAMC in Fig. 1 and document quantitative evaluations in Tab. 1 in terms of the average PSNR, RMSE, SSIM, and FID for (1) and (2) above; over four aforementioned datasets. For (1), we aim to have "better" scores, whereas all of these scores significantly getting worse means that diffusion outputs are very different for (2). Especially comparing FID scores, diffusion outputs of protected images are indeed adversarial for any model with no representative power.

		PSNR	RMSE	SSIM	FID
Wiki Art	(1)	25.98	7.90	0.87	123.52
	(2)	14.97	9.66	0.26	158.08
Historic Art	(1)	28.15	6.36	0.88	92.83
	(2)	16.24	9.42	0.32	163.80
Art 201	(1)	24.83	7.79	0.80	209.06
	(2)	15.73	9.68	0.29	241.43
Face Forensics	(1)	35.06	3.82	0.95	75.15
	(2)	22.40	8.81	0.73	106.96

Table 1. Quantitative similarities between (1) input and protected images and (2) diffusion outputs of them, over four datasets.

4.1. Comparison

We compare MAMC to image cloaking approaches such as [5, 23, 27, 28] with the same samples used in [28]. MAMC pushes the protected image to cause a significantly "bad" diffusion output. Note that, images created by other approaches are based on text-guidance like "A girl in the style of Karla Ortiz, black and white". In contrast, our target is to eliminate the image being used for guidance, so the diffusion output is as distorted as possible.Furthermore, if bad actors using diffusion models are not familiar with the artist's style, they may still distribute outputs created from

other cloaked images as in the artists' style without recognizing the difference, which is also damaging.

4.2. Protecting Artists from Style Infringement

Style transfer applications claim to have stolen artists' identity, as their style equates to their art. We evaluate MAMC on small single artist datasets to air its novelty. Fig. 4 samples the works of Edouard Manet and Francesco Albani, showcases how easy it is to replicate their style, and how MAMC protected versions do not allow that replication, along with the evaluation scores on the whole dataset.

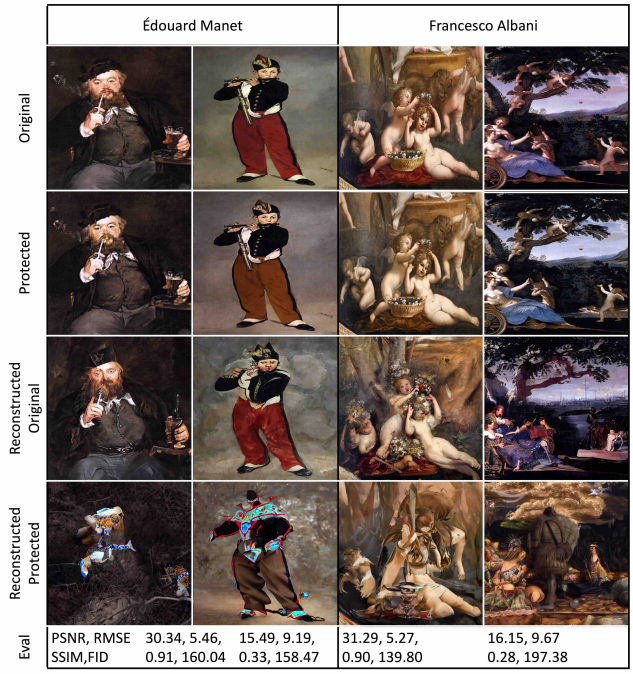

Figure 4. Diffusion models fail to replicate artists' style from adversarially protected images by MAMC.

4.3. Protecting Celebrities from Deepfakes

Another popular use case of diffusion models is personalization, which means fine-tuning the model on a specific face to create look-alikes. We test MAMC on a face dataset to verify that it can also be used proactively against potential deepfakes. Fig. 5 depicts two sample faces, their reconstructed, protected, and failed-to-reconstruct versions after MAMC, from full-datasets results in Tab. 1.

4.4. User Control

As mentioned, artists should have freedom over how much preservation and protection is applied by MAMC. In Fig. 6, five levels of the balance factor create varying changes upon the artwork. As expected, when 90% protection is desired (low α_R), protected image does not look like the input. Decreasing it to 80% uncovers some input image features. At 50% (similar α values), we can see the image preserved with the protection and the protection yielding a different diffusion output. At 10% protection (high α_R), changes are imperceptible with significantly different diffusion outputs.

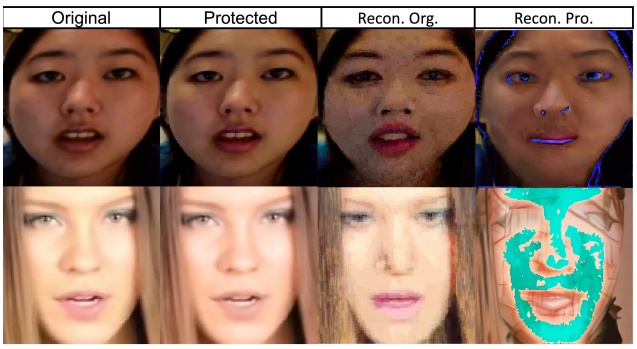

Figure 5. Diffusion models fail to replicate faces from adversarially protected images by MAMC.

Figure 6. The impact of the user balance variable in different levels. Percentage is the amount of protection (inverse fidelity).

5. Conclusion

We present "My Art My Choice", an adversarial protection model to prevent images from being exploited by diffusion models. There is a need for protection of copyrighted material and our cross-domain protector is ideal for interrupting diffusion-based tasks, such as personalization, style transfer, and any guided image-to-image translation. We evaluate MAMC on four datasets, assess user control, and compare to image cloaking (which is not adversarial protection). As generative AI services are proliferating, proactive protection services based on MAMC will also be valuable.

References

[1] Masaki Aono and Tosiyasu L Kunii. Botanical tree image generation. *IEEE computer graphics and applications*, 4(5): 10–34, 1984. 1

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023. 2

[3] Blake Brittain. Getty images lawsuit says stability ai misused photos to train ai. https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/, 2023. Accessed: 2023-08-10. 1

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. 2

[5] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*, 2021. 2, 3

[6] Umur A Ciftci, Gokturk Yuksek, and Ilke Demir. My face my choice: Privacy enhancing deepfakes for social media anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1379, 2023. 2

[7] Ilke Demir, Daniel G Aliaga, and Bedrich Benes. Proceduralization for editing 3d architectural models. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 194–202. IEEE, 2016. 1

[8] David S Ebert, Randall M Rohrer, Christopher D Shaw, Pradyut Panda, James M Kukla, and D Aaron Roberts. Procedural shape generation for multi-dimensional data visualization. *Computers & Graphics*, 24(3):375–384, 2000. 1

[9] Benj Edwards. Artists stage mass protest against ai-generated artwork on artstation. https://arstechnica.com/information-technology/2022/12/artstation-artists-stage-mass-protest-against-ai-generated-artwork/, 2023. Accessed: 2023-08-10. 1

[10] Iuri Frosio and Jan Kautz. The best defense is a good offense: Adversarial augmentation against adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4067–4076, 2023. 2

[11] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 2

[12] Aditya Golatkar, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Training data protection with compositional diffusion models. *arXiv preprint arXiv:2308.01937*, 2023. 2

[13] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Talecrafter: Interactive story visualization with multiple characters. *arXiv preprint arXiv:2305.18247*, 2023. 2

[14] Timothy B. Lee. Stable diffusion copyright lawsuits could be a legal earthquake for ai. https://arstechnica.com/tech-policy/2023/04/stable-diffusion-copyright-lawsuits-could-be-a-legal-earthquake-for-ai/, 2023. Accessed: 2023-08-10. 1

[15] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2023. 2

[16] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[19] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[21] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2

[22] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 3

[23] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023. 2, 3

[24] Sophie R. Saremsky, Umur A. Ciftci, Emily A. Greene, and Ilke Demir. Adversarial deepfake generation for detector

misclassification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2

[25] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 2

[26] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20595–20605, 2023. 2

[27] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pages 1589–1604, 2020. 3

[28] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023. 2, 3

[29] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2

[30] John M. Snyder. Chapter 6 - applying interval methods to geometric modeling. In *Generative Modeling for Computer Graphics and CAD*, pages 163–217. Academic Press, 1992. 1

[31] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. 2

[32] Hui Sun, Tianqing Zhu, Zhiqiu Zhang, Dawei Jin, Ping Xiong, and Wanlei Zhou. Adversarial attacks against deep generative models on data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 2

[33] Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–10, 2023. 2

[34] Kaggle user ansonnn. Historic art dataset. https://www.kaggle.com/datasets/ansonnnnn/historic-art, 2023. Accessed: 2023-08-10. 3

[35] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. *arXiv preprint arXiv:2303.15433*, 2023. 2

[36] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4563–4573, 2023. 2

[37] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2

[38] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18381–18391, 2023. 2

[39] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

[41] Xin Zhang, Jiaxian Guo, Paul Yoo, Yutaka Matsuo, and Yusuke Iwasawa. Paste, inpaint and harmonize via denoising: Subject-driven image editing with pre-trained diffusion model. *arXiv preprint arXiv:2306.07596*, 2023. 2

[42] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 2

[43] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N. Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6027–6037, 2023. 2