References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Bijan Afsari, Roberto Tron, and René Vidal. On the convergence of gradient descent for finding the riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260, 2013. 3, 5

[3] Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitris Kalatzis, and Søren Hauberg. Pulling back information geometry. *arXiv preprint arXiv:2106.05367*, 2021. 1, 3, 4

[4] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017. 1, 2, 3, 4, 5

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, Jan. 2023. 7

[6] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 2

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[9] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005. 3

[10] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A Closer Look at Fourier Spectrum Discrepancies for CNN-generated Images Detection, Mar. 2021. arXiv:2103.17195 [cs, eess]. 1

[11] Ricky T. Q. Chen, Jens Behrmann, David K Duvenaud, and Joern-Henrik Jacobsen. Residual Flows for Invertible Generative Modeling. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 7

[12] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 973–982, June 2023. 7, 8

[13] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024. 1

[14] D.C. Dowson and B.V. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 7

[15] R. Durall, Margret Keuper, and J. Keuper. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 7, 8

[16] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier Spectrum Discrepancies in Deep Network Generated Images. In *Advances in Neural Information Processing Systems*, volume 33, pages 3022–3032. Curran Associates, Inc., 2020. 2, 7, 8

[17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 7

[18] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. 3

[19] Sylvestre Gallot, Dominique Hulin, Jacques Lafontaine, Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. Riemannian metrics. *Riemannian Geometry*, pages 51–127, 2004. 2

[20] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 7

[21] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. DeepFake Detection by Analyzing Convolutional Traces. Apr. 2020. 1

[22] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *NIPS*, 2017. 7

[23] Søren Hauberg. Only bayes should learn a manifold (on the estimation of differential geometric structure from data). *arXiv preprint arXiv:1806.04994*, 2018. 1, 3, 5, 6

[24] Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficient-VDVAE: Less is more, Apr. 2022. 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[26] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 7

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 7

[28] Xianxu Hou, L. Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141, 2017. 7

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2018. 2, 6, 7

[30] Tero Karras, Miika Aittala, Samuli Laine, Erik Harkonen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 1, 7

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 2, 6, 7

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten,

Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, Seattle, WA, USA, June 2020. IEEE. 7

[33] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft Truncation: A Universal Training Technique of Score-based Diffusion Model for High Precision Score Estimation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 11201–11228. PMLR, June 2022. 7

[34] Durk P Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 7

[35] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014. 5

[36] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017. 7

[37] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 2, 6, 7

[38] Huiling Le. Estimation of riemannian barycentres. *LMS Journal of Computation and Mathematics*, 7:193–200, 2004. 5

[39] John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018. 1, 2

[40] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 6, 7, 12

[41] Xuezhe Ma and Eduard H. Hovy. Macow: Masked convolutional generative flow. In *NeurIPS*, 2019. 7

[42] Tambiama Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021. 1

[43] Jonathan H Manton. A globally convergent numerical algorithm for computing the centre of mass on compact lie groups. In *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.*, volume 3, pages 2211–2216. IEEE, 2004. 5

[44] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2017. 7

[45] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of GAN-Generated Fake Images over Social Networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389, Apr. 2018. 7, 8

[46] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. Incremental learning for the detection and classification of GAN-generated images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Dec. 2019. 8

[47] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE international conference on image processing (ICIP)*, pages 4584–4588. IEEE, 2019. 2, 7, 8

[48] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, Amit K. Roy-Chowdhury, and B. S. Manjunath. Detecting GAN generated Fake Images using Co-occurrence Matrices, Oct. 2019. 7, 8

[49] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022. 1

[50] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 2

[51] US Copyright Office. Copyright registration guidance: Works containing material generated by artificial intelligence., 2023. 1

[52] Stanislav Pidhorskyi, Donald A. Adjeroh, and Gianfranco Doretto. Adversarial Latent Autoencoders. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14092–14101, Seattle, WA, USA, June 2020. IEEE. 7

[53] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2016. 7

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 7

[56] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1

[57] Takashi Sakai. *Riemannian geometry*, volume 149. American Mathematical Soc., 1996. 2

[58] Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the Frequency Bias of Generative Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 18126–18136. Curran Associates, Inc., 2021. 1

[59] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. 1

[60] Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. Manifpt: Defining and analyzing fingerprints of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10791–10801, 2024. 1, 2, 3, 4, 6, 7, 8, 12

[61] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, Jan. 2023. 7

[62] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui

Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[63] Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679. Curran Associates, Inc., 2020. 7

[64] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based Generative Modeling in Latent Space. In *Advances in Neural Information Processing Systems*, volume 34, pages 11287–11302. Curran Associates, Inc., 2021. 7

[65] Sheng-Yu Wang, O. Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot… for Now. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 7, 8

[66] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models. In *International Conference on Learning Representations*, Feb. 2022. 7

[67] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. 2022. 7

[68] Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. *arXiv preprint arXiv:2401.12255*, 2024. 2

[69] Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7555–7565, 2019. 1, 2, 7, 8

[70] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. StyleSwin: Transformer-based GAN for High-resolution Image Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11294–11304, New Orleans, LA, USA, June 2022. IEEE. 7

Acknowledgements

This work was supported by the National Science Foundation (award 2318101), and a research grant from Sandia National Laboratories. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

A. Dataset Creation

ManiFPT [60] provides an extensive benchmark dataset for evaluating model attribution across a large array of GMs, spanning 4 different training datasets and all 4 main GM familys (GAN, VAE, Flow, Diffusion). However, what it is currently lacking is the inclusion of multimodal models such as vision-language models. To bridge this gap, and to evaluate model attribution methods on a wider variety of models, we created an extended benchmark dataset that includes SoTA vision-language GMs. In particular, we include 4 SoTA models (last row of Tab. 2) that can generate images given input text prompts: Flux.1-dev, Stable-Diffusion-3.5, Dall-E-3, and Openjourney. For all these models, we used pre-trained models that are available either on Huggingface or on public Github repositories.

A.1. Details on dataset creation

GM-CelebA dataset. To construct a dataset of faces that resemble images in CelebA [40], we use the text prompt of "a face of celebrity" to each of the vision-language models. For example, for Flux.1-dev model, we use the Huggingface's ‘diffuser‘ library to download the model weights, and used each pretrained model with default sampling configurations to generate 10k images with this prompt.

GM-CIFAR10 dataset. To generate images like the data in CIFAR10, we created a text prompt for each class in CIFAR10 (*i.e.*, airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck), as “an image of {cifar10-class}”. We then provided this prompt to each of the vision-language models we added in Tab. 2, and used each pretrained model with its default sampling configurations to generate a total of 10k images per prompt per CIFAR10 class.

B. Experiments on robustness of model fingerprints against post-processing

We evaluate the robustness of different model-attribution methods against two common post-processing perturbations: gaus- sian blurring (with increasing standard deviations) and JPEG compression (with decreasing quality factors). We train attribution methods on the training set, and apply these perturbations only at test time, to evaluate attribution accuracies under the two post-processing operations. We evaluate the test accuracies on all four datasets, using all 12 baseline methods and our methods. Figure 4 shows our results. Our methods (colored purple) consistently outperform all baselines, under both perturbation types and across all datasets.

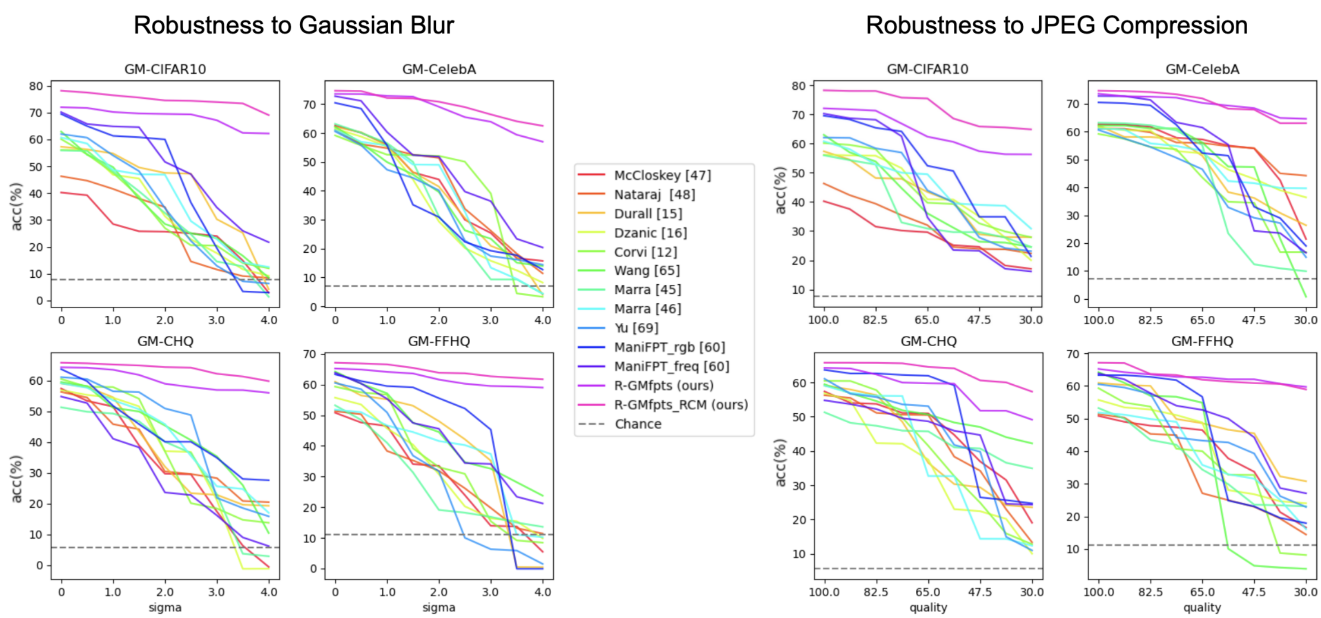

Figure 4. **Robustness** to test-time perturbation by Gaussian blurring (**Left**) and JPEG compression (**Right**). We evaluate the impact of two common test-time perturbations (Gaussian blurring and JPEG compression) on attribution accuracy across four GM datasets (GM-CIFAR10, GM-CelebA, GM-CHQ, GM-FFHQ). Each perturbation is applied only at the test-time. The plots report the accuracies of 13 attribution methods, including our proposed methods, R-GMfpts and R-GMfpts$_{\text{RCM}}$ (shown in purple). The dotted lines indicate chance-level accuracy. **(Left)** shows attribution accuracy as a function of sigma used in Gaussian blurs, and **(Right)** as a function of JPEG quality. Our methods consistently maintain higher attribution accuracy under both perturbations, demonstrating superior robustness across all four datasets. Link to larger image.