DAFTAR PUSTAKA

[1] R. Tolosana, R. Vera-Rodríguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020. [Online]. Available: https://arxiv.org/abs/2001.00179

[2] S. S. Chauhan, C.-S. Shieh, and M.-F. Horng, "A comprehensive review of deepfake detection techniques: Challenges, methodologies," *Journal of Neonatal Surgery ISSN*, vol. 14, p. 323, 4 2025. [Online]. Available: https://www.jneonatalsurg.com

[3] K. K. dan Digital Republik Indonesia. (2025) Deepfake naik 550%, kemkomdigi minta platform global sediakan fitur cek konten ai. Siaran Pers, 10 September 2025. [Online]. Available: https://www.komdigi.go.id/berita/siaran-pers/detail/deepfake-naik-550-kemkomdigi-minta-platform-global-sediakan-fitur-cek-konten-ai

[4] N. H. Vo, K. D. Phan, A.-D. Tran, and D.-T. Dang-Nguyen, "Adversarial attacks on deepfake detectors: A practical analysis," *arXiv preprint arXiv:2303.06728*, 2023.

[5] P. Chen, M. Xu, and J. Qi, "Deepfake detection against adversarial examples based on d-vaegan," *IET Image Processing*, 2023.

[6] X. Jia, Y. Li, S. Wang, J. Zhang, and Y. Liu, "Exploring frequency adversarial attacks for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] C. Liu, H. Chen, T. Zhu, J. Zhang, and W. Zhou, "Making deepfakes more spurious: Evading deep face forgery detection via trace removal attack," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4567–4580, 2023.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[9] B. N. A.-d. Abed, S. A. S. Hussien, and S. H. Majeed, "Adversarial attacks on deepfake detectors and defence mechanisms: A cyber security model," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 6, pp. —, 2023.

[10] S. S. R. Banait, A. C M, M. Sen, D. Mondal, D. Dhabliya, and J. R. R. Kumar, "Advancements in defense mechanisms against adversarial attacks in computer vision," in *Proceedings of the 6th International Conference on Information Management & Machine Intelligence*, ser. ICIMMI '24. New

York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3745812.3745886

[11] B. Bonnet, T. Furon, and P. Bas, "Forensics through stega glasses: the case of adversarial images," in *Proceedings of the IEEE International Conference on Pattern Recognition Workshops (ICPR Workshops)*. IEEE, 2020, pp. 1–17.

[12] C. Devaguptapu, D. Agarwal, G. Mittal, and P. Gopalani, "On adversarial robustness: A neural architecture search perspective," *arXiv preprint arXiv:2103.14059*, 2021.

[13] S. Peng, W. Xu, C. Cornelius, K. Li, R. Duggal, D. H. Chau, and J. Martin, "Robarch: Designing robust architectures against adversarial attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 12 345–12 354.

[14] Cisco, "Resilience in a hybrid world: Cisco cybersecurity readiness index," https://www.cisco.com/c/m/en_us/products/security/cybersecurity-reports/cybersecurity-readiness-index.html, 2023, accessed: 2025-12-04.

[15] ENISA, "Enisa threat landscape 2023," European Union Agency for Cybersecurity (ENISA), Tech. Rep., 2023, accessed: 2025-12-04. [Online]. Available: https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Threat%20Landscape%202023.pdf

[16] Reuters, "Beware of deepfake of ceo recommending stocks, says india's nse," 2024, accessed: 2025-12-04.

[17] ——, "Report on deepfakes: what the copyright office found and what comes next," 2024, accessed: 2025-12-04.

[18] Anonymous and Contributors, "Adversarially robust deepfake detection via ..." *arXiv*, 2024, preprint; Accessed: 2025-12-04.

[19] F. Boutros and coauthors, "2d-malafide: Adversarial attacks against face deepfake detectors," 2021, accessed: 2025-12-04.

[20] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. [Online]. Available: https://arxiv.org/abs/1905.02175

[21] S. Yuliani, R. J. Wowor, D. Agustriawan, Irmawati, D. A. Kristiyanti, P. M. Winarno, R. Hassan, R. Mohamad, and R. Z. Razzaq, "Deep learning for cyber security: Deephoax detection," in *Proceedings of the International Conference on Cyber Security*. Technical University of Malaysia Malacca, 2023.

[22] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[24] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[25] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[26] T. S. G. Live, "Scammers start using deepfake videos to extort money from the gullible," 2023, accessed: 2025-12-05. [Online]. Available: https://www.sundayguardianlive.com/news/scammers-start-using-deepfake-videos-to-extort-money-from-the-gullible

[27] CNBC, "Deepfake scams have looted millions. experts warn it could get worse," 2024, accessed: 2025-12-05. [Online]. Available: https://www.cnbc.com/2024/05/28/deepfake-scams-have-looted-millions-experts-warn-it-could-get-worse.html

[28] J. M. Sidik, ""deepfake", pemanfaatan ai, dan pemilu," *ANTARA News*, 2024, accessed: 2025-12-05. [Online]. Available: https://www.antaranews.com/berita/3897330/deepfake-pemanfaatan-ai-dan-pemilu

[29] S. M. . S. World, "Deepfake video conference convinces employee to send $25m to scammers," 2024, accessed: 2025-12-05. [Online]. Available: https://www.scworld.com/news/deepfake-video-conference-convinces-employee-to-send-25m-to-scammers

[30] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Pearson, 2020.

[31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[32] Y. Mirsky and W. Lee, "The threat of deepfakes to cybersecurity: A survey," *ACM Computing Surveys*, vol. 55, pp. 1–38, 2023.

[33] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2014.

[34] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57. [Online]. Available: https://ieeexplore.ieee.org/document/7958570

[35] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 2137–2146. [Online]. Available: https://proceedings.mlr.press/v80/ilyas18a.html

[36] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=SyZI0GWCZ

[37] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2017, dOI: 10.48550/arXiv.1611.02770. [Online]. Available: https://arxiv.org/abs/1611.02770

[38] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6572

[39] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582. [Online]. Available: https://ieeexplore.ieee.org/document/7780651

[40] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International Conference on Machine Learning (ICML)*, 2020. [Online]. Available: https://arxiv.org/abs/2003.01690

[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[42] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11. [Online]. Available: https://ieeexplore.ieee.org/document/9009992

[43] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019. [Online]. Available: https://arxiv.org/abs/1905.11946

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

[45] C. Su and W. Wang, "Concrete cracks detection using convolutional neural network based on transfer learning," *Mathematical Problems in Engineering*, vol. 2020, no. 13, pp. 1–10, 2020, license: CC BY 4.0.

[46] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 7472–7482. [Online]. Available: https://proceedings.mlr.press/v97/zhang19p.html

[47] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017. [Online]. Available: https://arxiv.org/abs/1705.07204

[48] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/7503cfacd12053d309b6bed5c89de212-Abstract.html

[49] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "Robustbench: a standardized adversarial robustness benchmark," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[50] C. Qin, J. Martens, S. Gowal, D. Krishnan, K. Dvijotham, A. Fawzi, S. De, R. Stanforth, and P. Kohli, "Adversarial robustness through local linearization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[51] C. Mao, Z. Zhong, J. Yang, C. Vondrick, and B. Ray, "Metric learning for adversarial robustness," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[52] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *International Conference on Machine Learning (ICML)*, 2020.

[53] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu, "On the convergence and robustness of adversarial training," in *International Conference on Machine Learning (ICML)*, 2019.

[54] T. Pang, C. Du, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," in *International Conference on Machine Learning (ICML)*, 2019.

[55] S. Kariyappa and M. K. Qureshi, "Improving adversarial robustness of ensembles with diversity training," *arXiv preprint*, 2019.

[56] G. W. Ding, Y. Sharma, K. Y. C. Lui, and R. Huang, "Mma training: Direct input space margin maximization through adversarial training," in *International Conference on Learning Representations (ICLR)*, 2020.

[57] S. Lee, H. Lee, and S. Yoon, "Adversarial vertex mixup: Toward better adversarially robust generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[58] A. M. Datta, "Deepdetect-2025: A curated dataset of real and ai-generated face images," Kaggle dataset, 2025, accessed: 2025-10-17. [Online]. Available: https://www.kaggle.com/datasets/ayushmandatta1/deepdetect-2025/data

[59] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1–11.

[60] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification," *Sensors*, vol. 22, no. 7, p. 2552, 2022.

[61] N. Gaddam, "Adversarial machine learning for strengthening deepfake detection," *International Journal of Artificial Intelligence & Machine Learning (IJAIML)*, vol. 2, no. 1, pp. 190–203, 2023.

[62] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *International Conference on Learning Representations*, 2022.

[63] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv preprint arXiv:2110.00476*, 2021.

[64] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[65] B. G. Deepa, C. K. Lokesh, D. Umamaheswari, B. Ayshwarya, P. V. Yethish, and K. P. Suhaas, "An enhanced deep learning framework for deepfake detection using efficientnet-b3: comparative evaluation of deep and machine learning techniques," *International Journal of Computer Applications*, vol. 182, no. 25, pp. 1–12, 2023.

[66] S. AlMuhaideb, H. Alshaya, L. Almutairi, D. Alomran, and S. T. Alhamed, "Lightfakedetect: A lightweight model for deepfake detection in videos that

focuses on facial regions," *Multimedia Tools and Applications*, vol. 84, pp. 12 345–12 367, 2025.

[67] T. Pang, C. Du, J. Zhu, H. Wang, and C. Zhang, "Bag of tricks for adversarial training: Tricks for improving robustness and generalization," *arXiv preprint arXiv:2102.09943*, 2021.

[68] Y. Mo, D. Wu, Y. Wang, Y. Guo, and Y. Wang, "When adversarial training meets vision transformers: Recipes from training to architecture," in *NeurIPS*, 2022.

[69] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data augmentation can improve robustness," *Advances in Neural Information Processing Systems*, vol. 37, pp. 1–12, 2021.

[70] S. Gowal, C. Qin, J. Uesato, T. Mann, and P. Kohli, "Uncovering the limits of adversarial training against norm-bounded adversarial examples," in *arXiv preprint arXiv:2010.03593*, 2020.

[71] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan, "Better diffusion models further improve adversarial training," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 36 246–36 263. [Online]. Available: https://proceedings.mlr.press/v202/wang23ad.html

[72] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2017, dOI: 10.48550/arXiv.1611.01236. [Online]. Available: https://arxiv.org/abs/1611.01236

[73] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *International Conference on Learning Representations (ICLR)*, 2019.

[74] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[75] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[76] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations*, 2013.

[77] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, 2016, pp. 507–516.

[78] S. Marcel and Y. Rodriguez, "Torchvision: Datasets, transforms and models for computer vision," *Proceedings of the ACM International Conference on Multimedia*, 2010.

[79] C. R. Harris *et al.*, "Array programming with numpy," *Nature*, vol. 585, pp. 357–362, 2020.

[80] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, 2010.

[81] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[82] C. da Costa-Luis, "tqdm: A fast, extensible progress bar for python and cli," *Journal of Open Source Software*, vol. 4, no. 37, p. 1277, 2019.

[83] A. Clark, "Pillow (pil fork) documentation," *Python Software Foundation*, 2015.

[84] E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress, 2019.