Layer Consistency Matters: Elegant Latent Transition Discrepancy for Generalizable Synthetic Image Detection

Supplementary Material

1. Overview

In this supplementary material, we first provide a comprehensive introduction to the datasets used in our experiments (Sec. 2). Then, we conduct additional experiments on the Chameleon dataset [25] (Sec. 3). More ablation studies related to visualization analysis, different degradation factors, and alternative backbones are presented (Sec. 4).

2. Datasets

UnivFD [14] contains synthetic images generated by a wide range of models including: 1) GAN-based methods. ProGAN [11], StyleGAN [12], BigGAN [3], CycleGAN [28], StarGAN [6], GauGAN [15], DeepFakes [18]; 2) Diffusion-based methods. ADM [7] trained on ImageNet, LDM [17], Glide [13], DALL-E [16]; and 3) The variants of some diffusion models. LDM with different noise refinement steps (*e.g.*, 100 vs. 200, with or without classifier-free guidance) and Glide with multi-stage refinement (*e.g.*, 100-27, 50-27, and 100-10). All samples are standardized to 256×256 resolution, with real counterparts sampled from the LAION [20] and ImageNet [19] datasets.

DRCT-2M [4] is a large-scale benchmark comprising 2 million diverse synthetic images generated by state-of-the-art diffusion models (DMs), including Stable Diffusion variants (v1.4, v1.5, SDXL) [17], ControlNet-enhanced models (*e.g.*, SDXL-Ctrl) [27], and high-speed inference variants (*e.g.*, SD-Turbo, LCM-SD). Images are sourced from community platforms (Civitai, Discord) to reflect real-world generation practices, capturing natural variations in prompt engineering, sampling configurations, and post-processing. Additionally, DRCT-2M contains a balanced subset of high-quality real photographs from established natural image collections to ensure comprehensive coverage of both authentic and synthetic content.

GenImage [29] is specifically designed for detecting synthetic images produced by modern generative models, which involves 1,331,167 real images and 1,350,000 fake images, carefully balanced in terms of class distribution and image count. The training set consists of images generated by Stable Diffusion v1.4 [17], paired with corresponding ImageNet [19] labels to ensure semantic alignment between real and synthetic data. During evaluation, the detector is tested on a diverse range of generators, including both diffusion models (*e.g.*, Stable Diffusion v1.5, GLIDE [13], ADM [7], VQDM [9], Wukong [2]) and GAN-based models (*e.g.*, BigGAN [3]), as well as commercial systems like Midjourney [1].

Chameleon [25] provides a high-quality and diverse collection of real and AI-generated images designed for evaluating detection robustness in realistic scenarios. It contains approximately 26K samples, with 14.8K real and 11.1K fake images, spanning four broad categories (human, animal, object, and scene), with resolutions ranging from 720P to 4K. The synthetic subset covers images generated from widely adopted text-to-image models such as Midjourney, Stable Diffusion (v1.4 and v1.5), DALLE-2, and various LoRA-based fine-tuned models. In Chameleon, the real subset is drawn from open-license platforms (*e.g.*, Unsplash) to ensure distributional alignment.

3. Comparisons on Chameleon

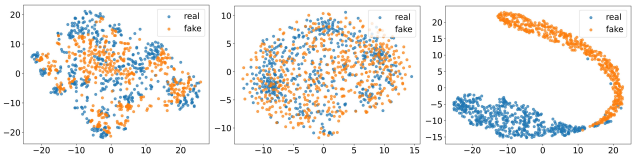
To further evaluate the robustness of our method in detecting realistic AI-generated content encountered in the wild, we conduct experiments on the Chameleon dataset, which contains diverse and high-quality samples that are inherently challenging to human perception. The results are reported in Table 1. We observe that all the models trained on ProGAN perform poorly, as the distributional gap between ProGAN-generated samples and the advanced generators used in Chameleon is substantial, resulting in near-random detection performance. In contrast, when training on more recent generators such as SDXL-Turbo and LCM-SDv1.5, the models achieve markedly improved performance, emphasizing the necessity of contemporary training data for generalizing to advanced synthetic media. Notably, our method consistently exceeds all baselines in both training datasets, demonstrating superior adaptability across generative distributions and validating its strong potential for real-world media forensics applications.

4. Ablation Study

Visualization of LTD. To demonstrate the effectiveness of our approach, we visualize the feature distributions using t-SNE across three different settings: (i) the ViT last-layer representation (Last Layer), (ii) the concatenated features from our selected discriminative layers (Selected Layers), and (iii) our LTD representations. As shown in figure 1, both the last-layer and selected layers remain heavily overlapped, exhibiting limited separability between real and synthetic images. In contrast, our full model LTD produces a clearly separable embedding space, indicating that LTD effectively exposes inter-layer discrepancies and bring out

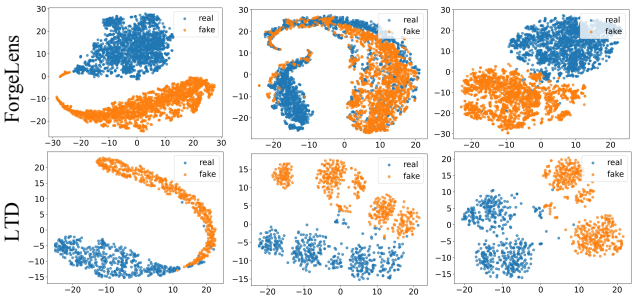
Table 1. Accuracy (Acc.) of different detectors (columns) on the Chameleon dataset when trained with different sources (rows). For each training dataset, the first row reports the overall classification accuracy, while the second row presents the class-wise accuracy split into "fake/real" for a more detailed breakdown.

Training Dataset	CNNSpot [24]	LGrad [22]	UnivFD [14]	NPR [23]	AIDE [25]	D^3 [8]	ForgeLens [5]	LTD (ours)
ProGAN	57.31 99.71/0.89	59.41 99.44/6.13	57.36 97.27/4.25	**59.83** 99.56/5.47	58.38 98.47/5.04	56.63 99.06/0.18	57.71 77.33/31.46	58.27 98.55/4.66
SDXL-Turbo	57.09 100.00/0.00	55.76 95.08/3.45	63.78 44.45/83.11	58.37 95.69/8.71	56.78 98.71/0.98	72.96 63.30/85.81	49.22 5.55/92.64	**73.98** 67.63/82.43
LCM-SDv1.5	61.96 96.16/16.45	53.30 73.81/26.00	68.00 54.94/81.06	60.28 87.98/23.41	62.60 86.66/30.68	70.50 72.99/67.18	57.09 100.00/0.00	**75.66** 80.60/60.09



(a) Last Layer (b) Selected Layers (c) LTD

Figure 1. t-SNE visualization of feature separability under different representations. Comparison among (a) the ViT last-layer features, (b) the concatenated selected mid-layer features, and (c) our LTD-enhanced representations. While both (a) and (b) exhibit limited separation between real and generated images, (c) shows clear and robust class separation, demonstrating the effectiveness of our LTD.



Original JPEG Downsampled

Figure 2. t-SNE visualization of feature distributions on SD v1.4 generated images. Columns show features for original images, (b) JPEG-compressed images, and downsampled iamges. Under JPEG compression, ForgeLens suffers from severe cluster collapse. Notably, under downsampling, while ForgeLens retains visual separability, it exhibits a significant distribution shift that is mismatched with with the classifier (resulting in 50% Acc).

discriminative artifacts that are otherwise not captured by the backbone alone.

Robustness Against Realistic Degradations. To evaluate the robustness of our method under realistic degradations, we conduct perturbation experiments on the Chameleon, using LCM-SDv1.5 as the training set. Specifically, we consider three widely encountered distortions: JPEG compression with a quality factor of 80, downsampling with a scale factor of 0.5, and Gaussian blur with a kernel size of 3. Each degradation is applied individually to isolate its impact, and jointly to simulate more challenging compound distortions. As shown in Table 2, LTD maintains stable performance under JPEG compression and downsampling, with only marginal decreases of 1.20% and 0.18% in Acc., respectively. Blur introduces more substantial perturbations, resulting in performance drops of 6.52% and 12.35%. When all distortions are combined, the accuracy decreases by 4.77% in Acc. and 7.37% in AP. Despite these challenges, the overall robustness of LTD across various perturbations underscores its ability to capture coarse-grained, degradation-tolerant artifacts that remain reliable even when fine-grained frequency cues are severely disrupted.

Table 2. Robustness analysis against image degradations on the Chameleon dataset. We evaluate the performance using average accuracy (Acc.) and average precision (AP). The values in parentheses denote the performance drop compared to the clean baseline.

JPEG	Downsampling	Blur	Acc.(%)	AP (%)
✗	✗	✗	75.66	78.00
✓	✗	✗	74.46 (1.20↓)	77.16 (0.84↓)
✗	✓	✗	75.48 (0.18↓)	77.41 (0.59↓)
✗	✗	✓	69.14 (6.52↓)	71.27 (6.73↓)
✓	✓	✓	70.89 (4.77↓)	70.63 (7.37↓)

To more intuitively show the discriminative ability of our LTD in dealing with realistic degradations, we visualize the t-SNE in feature space, where the most recent method ForgeLens [5] is employed as a reference. As shown in figure 2, in the ideal clean setting, both methods yield well-separated clusters between real and synthetic images. However, substantial differences emerge once degradations are

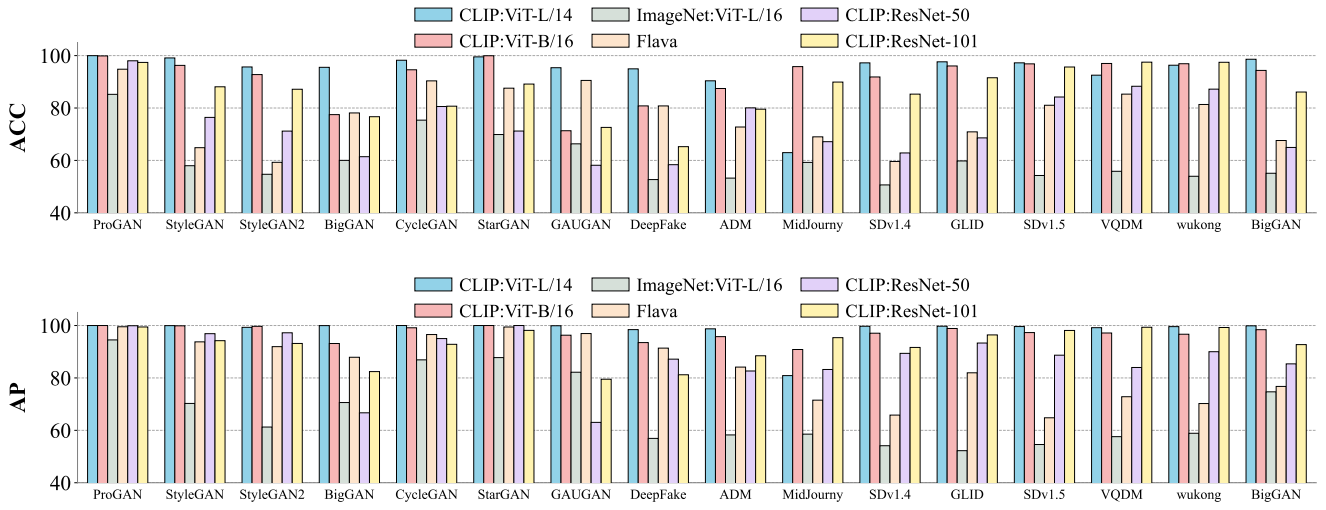


Figure 3. Backbone comparison across diverse generative models. Detection accuracy of four representative backbones, including CLIP ViT-L/14, CLIP ViT-B/16, ImageNet ViT-L/16, and Flava, evaluated over a wide range of GAN- and diffusion-based generators. The results highlight significant performance differences driven by pre-training strategy and model capacity.

applied. Under JPEG compression, LTD maintains a clear separation, while ForgeLens collapses into heavily overlapping clusters, showing almost no discriminative boundary. This behavior is consistent with the underlying mechanism of ForgeLens. ForgeLens is not a frequency-based detector, as it does not employ explicit spectral transformations or high-pass filtering. However, the method guides the frozen ViT to focus on forgery-specific local artifacts, which often reside in relatively high-frequency regions of the image. Consequently, when high-frequency content is suppressed by blur or corrupted by noise, these cues become less reliable, leading to the observed performance degradation.

For downsampling attacks, ForgeLens exhibits a significant performance discrepancy under downsampling attacks, where the detection accuracy drops to random guessing (as reported in the main manuscript) and the t-SNE visualization of the feature space reveals distinct, separable clusters. This mismatch indicates that downsampling does not fundamentally collapse the underlying feature distributions; instead, it disrupts the fine-grained local artifacts that ForgeLens relies on to form its decision boundary. As these high-frequency–sensitive cues are smoothed or removed, the classifier trained on them becomes invalid, even though coarser structural cues still induce separable feature clusters.

By comparison, LTD maintains clear separability under both JPEG compression and downsampling, because its cross-layer consistency cues operate at a coarse granularity that is inherently resistant to such degradations.

Impact of Different Backbones. To examine the influence of different backbones on AI-generated images detection, we conduct experiments on several representative architectures, including CLIP-based ViTs, ResNets, ImageNet-pretrained ViT, and Flava [21]. All backbones are trained on the 2-class training setting (*chair*, *tvmonitor*) described in the main paper, utilizing 72k ProGAN-generated images and real images from LSUN [26]. For the ResNet variants (CLIP:ResNet50 [10] and CLIP:ResNet101), since the feature dimensions vary across different stages, we apply Global Average Pooling to eliminate the spatial dimensions and align the features and project multi-stage features into a unified latent dimension. We summarize the performance across sixteen generators, spanning both GANs and DMs, as shown in Figure 3. the ImageNet-pretrained ViT-L/16 performs the worst, with accuracy dropping sharply across nearly all sources. This performance gap is primarily due to the significant difference in pre-training data scale; the limited volume of ImageNet-1K (1.28M images) prevents the model from developing the robust sensitivity to generative artifacts that emerges from CLIP's much larger dataset (400M image-text pairs). Similarly, Flava, although its multimodal training, remains inferior to CLIP, further confirming that pre-training data scale is a important factor. CLIP's extensive data provides a far superior prior for artifact sensitivity compared to the more limited datasets used for ImageNet or Flava. Furthermore, we observe a clear scaling effect across architecture, CLIP:ResNet-101 consistently outperforms CLIP:ResNet-50, and CLIP:ViT-L/14 exhibits greater stability than CLIP:ViT-B/16, which shows moderate but unstable performance on complex models like BigGAN and Stable Diffusion. These results prove that both increased model capacity and pre-training data scale are critical for developing a robust detector capable of handling the diverse artifacts of modern generative models.

Table 3. Ablation study on the selection endpoint (upper bound).

Layer Selection	UFD	DRCT-2M	GenImage	Mean Acc
Fixed(11-23)	95.99	91.26	91.70	92.98
Fixed(11-21)	94.47	87.09	94.55	92.04
Fixed(11-19)	95.11	91.63	97.44	94.73
Fixed(11-17)	97.03	91.11	94.72	94.29

Impact of Weight sharing To further investigate the efficacy of the weight-sharing mechanism within our dual-branch architecture, we conducted an additional ablation study comparing against a non-shared weight variant. In the non-shared setting, we assigned independent Transformer blocks to the selected raw feature branch and the LTD branch, respectively. The quantitative results demonstrate that the non-shared configuration leads to a consistent performance degradation across all evaluated benchmarks: on the UFD dataset , the accuracy dropped to 95.44%, which is 1.46% lower than the weight-sharing baseline; on the GenImage dataset , the performance decreased to 86.42%, representing a margin of 3.20% below the shared setting; and on the DRCT-2M dataset, the accuracy fell to 94.73%, showing a substantial drop of 4.81% compared to the proposed model. These results validate the critical role of weight sharing in our framework.

Impact of Weight sharing To further investigate the efficacy of the weight-sharing mechanism within our dual-branch architecture, we conducted an additional ablation study comparing against a non-shared weight variant. In the non-shared setting, we assigned independent Transformer blocks to the selected raw feature branch and the LTD branch, respectively. The quantitative results demonstrate that the non-shared configuration leads to a consistent performance degradation across all evaluated benchmarks: on the UFD dataset , the accuracy dropped to 95.44%, which is 1.46% lower than the weight-sharing baseline; on the GenImage dataset , the performance decreased to 86.42%, representing a margin of 3.20% below the shared setting; and on the DRCT-2M dataset, the accuracy fell to 94.73%, showing a substantial drop of 4.81% compared to the proposed model. These results validate the critical role of weight sharing in our framework.

Impact of Upper Bound To further refine the optimal search space, we fixed the starting layer at Layer 11 (the optimal lower bound determined in our main text) and varied the endpoint from Layer 17 to Layer 23. As demonstrated in Table 3, the detection performance consistently improves as the endpoint shifts from 23 towards 19. While extending the boundary to 17 yields the highest accuracy on the UFD dataset (97.03%), it results in a significant performance drop on DRCT-2M (94.29%) compared to the Layer 19 configuration. This indicates that a narrower selection at Layer 17 leads to overfitting and limited generalization across diverse datasets.

References

[1] Midjourney. In https://www.midjourney.com/home/, 2022. 1

[2] Wukong. In https://xihe.mindspore.cn/modelzoo/wukong, 2022. 1

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 1

[4] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *International Conference on Machine Learning*, 2024. 1

[5] Yingjian Chen, Lei Zhang, and Yakun Niu. Forgelens: Data-efficient forgery focus for generalizable forgery image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16270–16280, 2025. 2

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794, 2021. 1

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633, 2024. 2

[9] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1

[10] 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 (CVPR)*, pages 770–778, 2016. 3

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[13] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion

models. In *International Conference on Machine Learning*, pages 16784–16804, 2022. 1

[14] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 1, 2

[15] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[16] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831, 2021. 1

[17] 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. 1

[18] 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 (ICCV)*, 2019. 1

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[20] 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. In *Data Centric AI NeurIPS Workshop 2021*, 2021. 1

[21] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[22] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 2

[23] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2

[24] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 2

[25] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, 2025. 1, 2

[26] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 3

[27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. 1

[29] Mingjian Zhu, Hanting Chen, Qiangyu YAN, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *Advances in Neural Information Processing Systems*, pages 77771–77782, 2023. 1