ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095167

ON THE DETECTION OF SYNTHETIC IMAGES GENERATED BY DIFFUSION MODELS

Riccardo Corvi[⋆]*, Davide Cozzolino* [⋆]*, Giada Zingarini*[⋆]*, Giovanni Poggi*[⋆]*, Koki Nagano*[†]*, Luisa Verdoliva*[⋆]

⋆ University Federico II of Naples, Italy
† NVIDIA

ABSTRACT

Over the past decade, there has been tremendous progress in creating synthetic media, mainly thanks to the development of powerful methods based on generative adversarial networks (GAN). Very recently, methods based on diffusion models (DM) have been gaining the spotlight. In addition to providing an impressive level of photorealism, they enable the creation of text-based visual content, opening up new and exciting opportunities in many different application fields, from arts to video games. On the other hand, this property is an additional asset in the hands of malicious users, who can generate and distribute fake media perfectly adapted to their attacks, posing new challenges to the media forensic community. With this work, we seek to understand how difficult it is to distinguish synthetic images generated by diffusion models from pristine ones and whether current state-of-the-art detectors are suitable for the task. To this end, first we expose the forensics traces left by diffusion models, then study how current detectors, developed for GAN-generated images, perform on these new synthetic images, especially in challenging social-network scenarios involving image compression and resizing. Datasets and code are available at https: `github.com/grip-unina/DMimageDetection`.

Index Terms— Synthetic image detection, GANs, Diffusion models, Text-to-image.

1. INTRODUCTION

The use of diffusion models for the generation of synthetic media is arousing great interest among both researchers and practitioners. Besides the high quality and photorealism of the generated images, it is the opportunity to model a wide variety of subjects and contexts that appears to be particularly interesting. In fact, diffusion models can be guided by textual descriptions or pilot sketches to generate images of a virtually unlimited set of categories, bounded only by our imagination. Therefore, this technology represents a powerful tool for artists, game designers, and any type of creative users. Unfortunately, this includes also malicious users, that may take advantage of this increased flexibility to generate fake media more fit to their disinformation goals. In this paper, we try to assess how prepared we are to face this new threat.

Fig. 1: Synthetic images generated using recent text-to-image models: DALL·E 2 [3], stable diffusion [4] and GLIDE [5].

Some recent papers have begun studying the detection of DM-based images. In particular, in [1, 2] it was noted that the lack of explicit 3D modeling of objects and surfaces causes asymmetries in shadows and reflected images. Furthermore, global semantic inconsistency can be observed, to some extent, in lighting. These traces can certainly be exploited to identify today's DM images. However, if the rapid advancement of GAN images can be taken as a paradigm, new DM-based methods can be expected to soon overcome these limitations and generate images that satisfy all necessary semantic constraints, be it lighting, perspective or any other type.

Indeed, most state-of-the-art detectors rely on traces that are invisible to the human eye. Even images that are visually perfect, with no evident semantic inconsistency, can be distinguished from real images based on the traces that are inherent of the generation process. In fact, any method used to create synthetic visual data embeds some peculiar traces in their

This material is based on research sponsored by DARPA under agreement number FA8750-20-2-1004. In addition, this work has received funding by the European Union under the Horizon Europe vera.ai project, Grant Agreement number 101070093, and is supported by a TUM-IAS Hans Fischer Senior Fellowship and the PREMIER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program.

output images, that are related to the actions taken in the generation process. These traces are different from those typical of modern digital devices enabling fake image detection [6]. Moreover, each generation architecture is characterized by its own peculiar traces, thereby allowing also for source attribution. The presence and distinctiveness of such traces have been proven by extracting a sort of spatial-domain artificial fingerprints [7, 8], but also through frequency-domain analyses showing that the upsampling operation performed in most GAN architectures give rise to clear spectral peaks [9, 10].

Based on these concepts, several promising CNN-based detectors of GAN images have been developed. However, the problem is far from being solved. On one hand, new and more sophisticated generation architectures are proposed by the day, some of them, like StyleGAN3 [11], aiming explicitly at minimizing the presence of these undesired traces. On the other hand, even state-of-the-art detectors, based on a supervised learning approach, have a hard time generalizing to architectures never seen in training. Moreover, they suffer a significant performance drop when image quality is impaired, as it always happens on social networks, which routinely apply some resizing and compression operations. This is because forensics traces are very weak and can be easily removed even by these non-malicious processing steps.

An interesting experiment on generalization has been recently conducted by NVIDIA within the DARPA SemaFor program. Performers were asked to detect StyleGAN3 images, with the constraint that no images generated with this architecture could be used in training. Despite the inherent difficulty of the task, good results have been achieved [12]. A more recent contest (IEEE VIP Cup) extended this analysis to images coming also from diffusion models architectures, with encouraging results [13]. Also in [14] an initial study on the detection and attribution of diffusion models is proposed, with promising outcomes. However, results are presented only in ideal conditions and no robustness analysis is considered.

This work aims at providing some more information on the detection of DM images and, possibly, contributing some useful guidelines for further developments. In particular, we want to answer two fundamental questions: *i)* are DM images characterized by hidden artifacts similar to those observed in GAN images? *ii)* To what extent are current state-of-the-art detectors effective on this type of images? To answer these questions, we generated a large variety of synthetic images using the most recent generators. Then, we carried out an analysis of their artifacts, and finally studied the performance of some deep learning-based detectors on them, not only in ideal conditions but also in more challenging scenarios, where images are compressed and resized.

2. BACKGROUND

In this section, we summarize the most important findings in the literature towards the development of successful and robust deep learning-based detectors for synthetic images. Much of the previous research relates to images generated by GANs as these architectures have so far dominated the field.

A widely agreed fact is the key importance of augmentation, including especially blurring and compression, to ensure robustness. On the same line, training set diversity was found to help generalizing to unseen architectures, as shown in [15] where a simple pre-trained Resnet50 is trained on 20 different categories of ProGAN images. Such observations have been reinforced by later works [16, 17] proving the tight relation between augmentation and training diversity, on one side, and detectors' reliability, on the other. Working on local patches also appears to be important [18] as well as analyzing both local and global features [19].

Another central discovery concerns the need to avoid any loss of information in the pre-processing of training and test images, as well as in all layers of the neural network, especially those closest to the input. Most of all, it is important to avoid resizing, a common practice in deep learning to adapt images to fixed input layers, as this entails image resampling and interpolation, which may erase the subtle high-frequency traces left by the generation process. To preserve these precious (and invisible) forensics artifacts, several strategies can be considered: *i)* training the networks on local patches, cropped from the image with no resizing; *ii)* making the final decision on the whole image through some fusion strategy; *iii)* avoid downsampling steps in the first layers of the network [16].

These measures allow to minimize information losses in the precious high-frequency image components. A more extensive analysis of the performance of synthetic image detectors [16] shows that pre-training all models on large datasets (e.g., ImageNet) keeps being important, while using residuals instead of the original images does not improve performance, and extreme augmentation provides only marginal gains.

3. ARTIFACT ANALYSIS

Previous work [7, 8] established the existence of GAN fingerprints and their dependence on both the GAN architecture (number and type of layers) and its specific parameters (filter weights). In particular, in [7], fingerprints are extracted in the spatial domain mimicking the pipeline used to extract the PRNU pattern for device identification [27]. First, the scene content is estimated through a denoising filter $f(\cdot)$, $\widehat{X}_i = f(X_i)$, and removed from the input image to obtain the so-called noise residual, $R_i = X_i - f(X_i)$. This latter is assumed to be the sum of a deterministic component, the GAN fingerprint F, and a random noise component W_i, so the fingerprint is estimated by simply averaging a large number of image residuals $\widehat{F} = (1/N) \sum_{i=1}^{N} R_i$.

In this work we use the same procedure and rely on the denoising filter proposed in [28], which proved successful for camera fingerprint extraction [29]. We average the noise

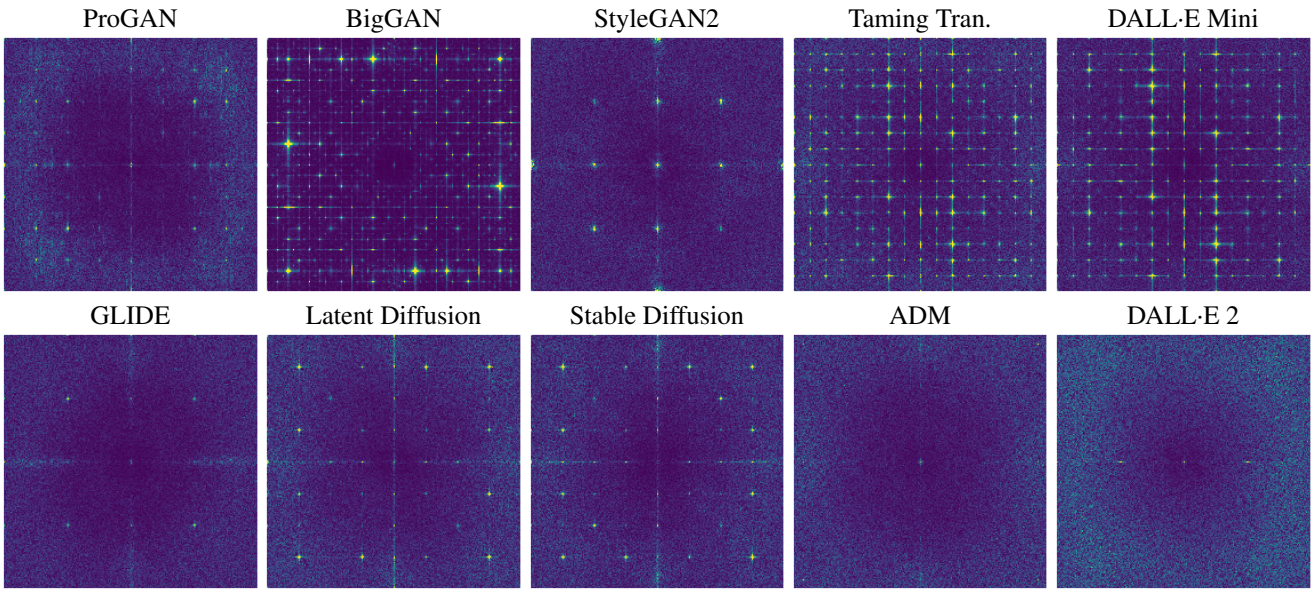

Fig. 2: Fourier transform (amplitude) of the artificial fingerprint estimated from 1000 image residuals. Top row: from left to right ProGAN [20], BigGan [21], StyleGAN2 [22], Taming Transformers [23], DALL·E Mini [24]. Bottom row: GLIDE [5], Latent Diffusion [25], Stable Diffusion [4], ADM [26], DALL·E 2 [3]

.

residuals of 1000 images, then take the Fourier transform of the result to carry out a spectral analysis. Fig.2 shows the amplitude of these spectra for several architectures of interest, both GANs [20, 21, 22] or VQ-GANs [23, 24] (top row), and DMs [5, 25, 4, 3] (bottom row). For all GANs, strong peaks are clearly visible in the spectra [9, 15], implying the presence of quasi-periodical patterns in the synthetic images. Interestingly, the same happens with some recent diffusion models, such as GLIDE, Latent Diffusion and Stable Diffusion, leading to expect good results for fingerprint-based forensic tools also in these cases. On the other hand, such peaks are much weaker for other architectures, like ADM and DALL·E 2, predicting more controversial results in these cases, as will be confirmed by the experimental analysis in the next Section.

4. DETECTION PERFORMANCE

In this Section we present the results of experiments carried out on images generated by several state-of-the-art generative models including GANs, transformers, and DMs: ProGAN [20], StyleGAN2 [22], StyleGAN3 [11], BigGAN [21], EG3D [30], Taming Transformer [23], DALL·E Mini [24], DALL·E 2 [3], GLIDE [5], Latent Diffusion [25], Stable Diffusion [4] and ADM (Ablated Diffusion Model) [26]. For text-to-image data we use language prompts from COCO validation and training set. Real data instead come from COCO [31], ImageNet [32] and UCID [33].

We are especially interested in the ability of detectors to generalize to unseen architectures, relevant for realistic scenarios. Therefore, we train on images generated by a single model, and test on all others. In detail, we consider two cases: a) training only on ProGAN images, using the same setting as in [15] (362K fake images with 20 categories); b) training only on images generated by Latent Diffusion (200K fake images with 5 categories). In both cases, tests are performed on 1000 synthetic images for each model and 5000 real images.

We compare the following detectors: Spec [9], based on frequency analysis, PatchForensics [18], that relies on local patch analysis, Wang2020 [15], a ResNet50 with blurring and compression augmentation, and Grag2021 [16], same backbone as before but including intense augmentation and avoiding down-sampling in the first layer. Results are given in terms of area under the receiver-operating curve (AUC) and accuracy at the fixed threshold of 0.5.

Generalization and robustness. First of all, we analyze the performance on uncompressed synthetic images in the PNG format, as originated from each model (Table 1, left). This first experiment highlights that in this case detection is very easy, because real images, which are always JPEG compressed by a codec embedded in the camera, are characterized by JPEG compression artifacts, while synthetic images do not embed such traces. In fact, the AUC performance is almost perfect on ProGAN (seen in training) and remains pretty good also for other architectures. Even in this favorable case, the accuracy is often unsatisfactory, because the threshold does not work well on images of different origin [16].

This is a simple scenario, compared to the situation where both synthetic and real images are compressed and resized

	Trained on ProGAN							
	Uncompressed				Resized and Compressed			
Acc./AUC%	Spec	PatchFor.	Wang2020	Grag2021	Spec	PatchFor.	Wang2020	Grag2021
ProGAN	83.5/ 99.2	64.9/ 97.6	99.9/100	99.9/100	49.7/ 48.5	50.4/ 65.3	99.7/100	99.9/100
StyleGAN2	65.3/ 72.0	50.2/ 88.3	74.0/ 97.3	98.1/ 99.9	51.8/ 50.5	50.8/ 73.6	54.8/ 85.0	63.3/ 94.8
StyleGAN3	33.8/ 4.4	50.0/ 91.8	58.3/ 95.1	91.2/ 99.5	52.9/ 51.9	50.2/ 76.7	54.3/ 86.4	58.3/ 94.4
BigGAN	73.3/ 80.5	52.5/ 85.7	66.3/ 94.4	95.6/ 99.1	52.1/ 52.2	50.5/ 58.8	55.4/ 85.9	79.0/ 99.1
EG3D	80.3/ 89.6	50.0/ 78.4	59.2/ 96.7	99.4/100	58.9/ 60.6	49.8/ 81.9	52.1/ 85.1	56.8/ 96.6
Taming Tran.	79.6/ 86.6	50.5/ 69.4	51.2/ 66.5	73.5/ 96.6	49.0/ 49.1	50.0/ 64.1	50.5/ 71.0	56.2/ 94.3
DALL·E Mini	80.1/ 88.1	51.5/ 82.2	51.7/ 60.6	70.4/ 95.6	59.1/ 61.9	50.1/ 68.7	51.1/ 66.2	62.3/ 95.4
DALL·E 2	82.0/ 93.0	50.0/ 51.1	50.3/ 85.8	51.9/ 94.9	61.8/ 64.5	49.8/ 58.3	49.9/ 46.1	50.0/ 65.4
GLIDE	73.4/ 81.9	50.3/ 96.6	51.1/ 62.6	58.6/ 86.4	53.1/ 52.5	51.0/ 71.5	50.3/ 65.9	51.8/ 90.0
Latent Diff.	72.1/ 78.5	51.8/ 84.3	51.0/ 62.5	58.2/ 91.5	47.9/ 46.3	50.6/ 65.2	50.7/ 69.1	52.4/ 89.4
Stable Diff.	66.8/ 74.7	50.8/ 85.0	50.9/ 65.9	62.1/ 92.9	46.5/ 44.5	51.1/ 77.2	50.7/ 72.9	58.1/ 93.7
ADM	55.1/ 53.3	50.4/ 87.1	50.6/ 56.3	51.2/ 57.4	49.1/ 49.1	51.0/ 69.1	50.3/ 68.1	50.6/ 77.2
AVG	70.5/ 75.2	51.9/ 83.2	59.5/ 78.6	75.8/ 92.8	52.7/ 52.7	50.4/ 69.2	55.8/ 75.1	61.5/ 90.8

Table 1: Comparative analysis of state-of-the-art techniques. All methods were trained only on ProGAN images, and tested both on uncompressed synthetic data (left) and on resized and compressed data (right).

as routinely happens on social media platforms. To simulate such forms of image laundering and to avoid polarization, we follow the procedure used in the IEEE VIP Cup [13]. For each image of the test, a crop with random (large) size and position is selected, resized to 200×200 pixels, and compressed using a random JPEG quality factor from 65 to 100. In this challenging condition, a general reduction of the performance is observed, Table 1 (right), except for ProGAN (present in training). Again, the performance is acceptable in terms of AUC, but almost random in terms of accuracy. The most difficult diffusion models are DALL·E 2 and ADM, which presented very weak artifacts in our previous analysis.

Then, we trained the best performing approach (Grag2021) on images generated by Latent Diffusion, testing it on the resized/compressed dataset. First, we observe (Table 2, left) that an almost perfect detection is achieved not only on Latent Diffusion but also on Stable Diffusion, coherently with the fact that these architectures share similar artifacts (Fig.2). The performance on other diffusion models, instead, are close to those obtained on GAN generated images. This means that Stable and Latent diffusion models are characterized by different cues compared to ADM and DALL·E 2.

Fusion and calibration. Finally, we carried out a simple experiment where we fuse (simple average) the outputs of the networks trained on both datasets. Results are reported again in Table 2. Of course, the performance on GAN generated images improves, while remaining reasonably good on diffusion models, however accuracy is still extremely low, due to the unsuitable fixed threshold. To improve accuracy we try using a calibration procedure (Platt scaling method [34]) under the hypothesis of having only 20 real images and 20 synthetic ones for each model. Performance greatly improves, but we still cannot reliably detect images that present artifacts significantly different from those seen during training.

Acc./AUC%	Trained on Latent Diffusion	Fusion	Calibration
ProGAN	52.0/ 78.3	90.2/100	99.6/100
StyleGAN2	58.0/ 85.0	56.6/ 94.6	85.8/ 94.6
StyleGAN3	59.5/ 87.6	55.4/ 93.9	85.4/ 93.9
BigGAN	52.9/ 80.6	59.3/ 98.5	92.7/ 98.5
EG3D	65.4/ 91.8	54.4/ 97.7	89.6/ 97.7
Taming Tran.	78.2/ 97.3	61.5/ 98.2	92.4/ 98.2
DALL·E Mini	73.9/ 97.3	65.9/ 97.7	90.3/ 97.7
DALL·E 2	50.3/ 73.3	50.0/ 72.2	50.0/ 72.2
GLIDE	62.5/ 96.2	52.5/ 95.9	88.6/ 95.9
Latent Diff.	97.1/ 99.9	84.9/ 99.8	97.6/ 99.8
Stable Diff.	99.7/100	92.5/100	99.5/100
ADM	52.9/ 81.9	50.8/ 80.6	72.1/ 80.6
AVG	66.9/ 89.2	64.5/ 94.1	87.0/ 94.1

Table 2: Results of Grag2021: trained only on Latent Diffusion (left), fused with Grag2021 trained only on ProGAN (center), with calibration applied after fusion (right).

5. CONCLUSION

This work addressed the problem of detecting synthetic images generated by diffusion models. We first tested whether DM images are characterized by distinctive fingerprints just as GAN images are, obtaining a partial confirmation. Then we analyzed the performance of some state-of-the-art detectors under different realistic scenarios. Experimental results vary significantly from model to model, as these are characterized by different forensics cues. Generalization remains the main hurdle, and detectors trained only on GAN images work poorly on these new generators. Including a DM model in training can help to detect images generated by similar diffusion models but results can be unsatisfactory for others. Of course, these are only preliminary results and deeper analyses are necessary to address the problem of DM image detection.

6. REFERENCES

[1] H. Farid, "Lighting (in)consistency of paint by text," *arXiv preprint arXiv:2207.13744v2*, 2022.

[2] H. Farid, "Perspective (in)consistency of paint by text," *arXiv preprint arXiv:2206.14617v1*, 2022.

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125v1*, 2022.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "Stable diffusion," `https://github.com/CompVis/stable-diffusion`, 2022.

[5] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

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

[7] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs Leave Artificial Fingerprints?," in *IEEE MIPR*, 2019, pp. 506–511.

[8] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in *ICCV*, 2019.

[9] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and Simulating Artifacts in GAN Fake Images," in *IEEE WIFS*, 2019, pp. 1–6.

[10] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," in *CVPR*, 2020.

[11] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *NeurIPS*, vol. 34, pp. 852–863, 2021.

[12] K. Nagano, "StyleGAN3 Synthetic Image Detection," `https://github.com/NVlabs/stylegan3-detector`, 2021.

[13] R Corvi, D. Cozzolino, K. Nagano, and L. Verdoliva, "IEEE Video and Image Processing Cup," `https://grip-unina.github.io/vipcup2022/`, 2022.

[14] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models," *arXiv preprint arXiv:2210.06998*, 2022.

[15] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *CVPR*, 2020.

[16] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art," in *IEEE ICME*, 2021.

[17] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro, "Detecting GAN-generated Images by Orthogonal Training of Multiple CNNs," in *IEEE ICIP*, 2022.

[18] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *ECCV*, 2020.

[19] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu, "Fusing Global and Local Features for Generalized AI-Synthesized Image Detection," in *IEEE ICIP*, 2022.

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in *ICLR*, 2018.

[21] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," in *ICLR*, 2018.

[22] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *CVPR*, 2020, pp. 8110–8119.

[23] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021, pp. 12873–12883.

[24] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Lê Khàc, L. Melas, and R. Ghosh, "DALL-E Mini," 7 2021.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[27] J. Lukàš, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, 2006.

[28] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[29] D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, no. 1, pp. 14–27, 2020.

[30] E. R. Chan, C. Z Lin, M. A Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3d generative adversarial networks," in *CVPR*, 2022, pp. 16123–16133.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[33] G. Schaefer and M. Stich, "UCID: An uncompressed color image database," in *Storage and Retrieval Methods and Applications for Multimedia*, 2003, vol. 5307, pp. 472–480.

[34] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," *Advances in Large Margin Classiers*, 1999.