FAKE OR JPEG? REVEALING COMMON BIASES IN GENERATED IMAGE DETECTION DATASETS

A PREPRINT

Patrick Grommelt[1,2,*], **Louis Weiss**[1,2,*], **Franz-Josef Pfreundt**[1], and **Janis Keuper**[2,3]

[1]Fraunhofer ITWM, Competence Center High Performance Computing Kaiserslautern, Germany
[2]Institute of Machine Learning and Analysis (IMLA), Offenburg University, Germany
[3]University of Mannheim, Germany

[*]*Equal contribution*

March 29, 2024

ABSTRACT

The widespread adoption of generative image models has highlighted the urgent need to detect artificial content, which is a crucial step in combating widespread manipulation and misinformation. Consequently, numerous detectors and associated datasets have emerged. However, many of these datasets inadvertently introduce undesirable biases, thereby impacting the effectiveness and evaluation of detectors.

In this paper, we emphasize that many datasets for AI-generated image detection contain biases related to *JPEG* compression and image size. Using the *GenImage* dataset, we demonstrate that detectors indeed learn from these undesired factors. Furthermore, we show that removing the named biases substantially increases robustness to *JPEG* compression and significantly alters the cross-generator performance of evaluated detectors. Specifically, it leads to more than 11 percentage points increase in cross-generator performance for *ResNet50* and *Swin-T* detectors on the *GenImage* dataset, achieving state-of-the-art results.

We provide the dataset and source codes of this paper on the anonymous website:
`https://www.unbiased-genimage.org`

Keywords Generation Detection · Diffusion Model · Bias · Dataset

1 Introduction

In recent years, generative models have improved significantly in creating photorealistic images, making it hard for humans to distinguish between natural and generated images [1]. Even though these advancements are a significant achievement in many computer vision applications, it does not take much to realize how the misuse of these models can threaten public safety. To tackle this issue, extensive research has been conducted on the development of robust detection mechanisms and techniques for identifying manipulated or synthetic content that is not actively fingerprinted by the generator side.

Generative Adversarial Networks (*GANs*) [2] have marked a major breakthrough in generating realistic synthetic images. Nevertheless, spectral analysis reveals that images produced by *GANs* inherently contain distinctive artifacts, making them recognizable. Zhang *et al.* [3] observed that *GAN*-generated images exhibit periodic, grid-like patterns in their frequency spectrum, a clear deviation from natural image spectra. Furthermore, Durall *et al.* [4] pointed out that *GANs* fail to replicate the spectral distribution of training data, probably due to the inherent transposed convolution operations of these models. This discrepancy between natural and generated images has enabled the research community to develop highly effective, generator-agnostic detection tools, achieving near-perfect accuracy in identifying synthetic images.

However, the landscape of synthetic content generation has evolved rapidly. One significant advancement came with the introduction of *Denoising Diffusion Probabilistic Models* (*DDPM*) [5], which represented a paradigm shift in generative modelling. *DDPMs* employ a fundamentally different approach compared to *GANs*, utilizing a diffusion process to generate images, which results in a smoother and more realistic appearance. These generative models have also shown better capabilities in approximating the frequency spectra of natural images [6, 7, 8], which caused a lot of detectors developed for the identification of *GAN* generated images to perform poorly on these images [6]. To the best of our knowledge, the problem of finding generator-agnostic and robust detection methods for *DDPM*-generated images has not been solved yet. Even though it seems trivial to correctly classify images which are generated by the same model the detector has seen in training, it does not transfer well to other generative models [6, 7, 9, 10].

To effectively assess and compare the generalization capabilities of detectors, it is imperative to establish a consensus within the community regarding the benchmark to be employed. Ideally, such benchmarks should closely mimic real-world scenarios by being large-scale and encompassing a diverse array of classes and images produced by a range of distinct generators. To address this need, *GenImage* [9] introduced an extensive dataset that includes all natural images from the *ImageNet1k* dataset [11], as well as approximately an equal number of generated images, originating from various generators with differing architectures. Nevertheless, the evaluation of a detector based on the raw *GenImage* Benchmark is not reliable yet since, as we will show, *JPEG* compression and image size biases are used by detectors during the training.

This paper's primary goal is to raise awareness of *JPEG* compression and image size biases present in most datasets for generated image detection and emphasize the need for heightened scrutiny to ensure that detectors do not inadvertently learn from undesirable variables. Current evaluation benchmarks are somewhat limited in their interpretability, because they do not ensure whether the detection is really based on generation-specific artifacts and thus applicable for real world usage. Consequently, it becomes challenging to assess the effectiveness of various approaches and determine which ideas warrant further research. We firmly believe that identifying and mitigating biases in datasets for generated image detection is crucial, as it lays the foundation for establishing a robust and transparent research environment for generated image detection.

In summary, our main contributions are as follows:

• We demonstrate, using the *GenImage* dataset as an example, that many datasets for generated image detection contain *JPEG* and image size biases, which are subsequently used by the detectors during inference.

• We show that removing these biases significantly enhances cross-generator performance, achieving state-of-the-art results on *GenImage* and increasing the average accuracy by more than 11 percentage points for *ResNet50* and *Swin-T* detectors. Additionally, detectors become more robust against distortions due to now learning the actual task of detecting generation specific artifacts.

2 Related Works

2.1 Detecting Artifically Generated Images

The problem of detecting AI-generated images is traditionally approached as a binary classification task [4, 3, 12, 13, 14]. However, for practical use in real-world scenarios, detector models not only need to perform well on the training and validation datasets but also exhibit two critical characteristics. Firstly, they must generalize effectively to generative models that were not part of their training data, as the specific generator is typically unknown. This concept is referred to as cross-generator performance. Secondly, these models need to be robust against various transformations such as resizing, compression, or noise, as attackers may attempt to manipulate content to deceive the detector and social media platforms often apply these transformations by default.

Detecting GAN Generated Images

Research has indicated that *GAN*-generated images contain distinctive patterns in the frequency spectrum of their Discrete Cosine Transformation [3]. These patterns arise due to the presence of upsampling layers [4], making *GAN*-generated content detectable. Wang *et al.* [12] have demonstrated that even a basic CNN classifier, *ResNet50* [15], can successfully identify images generated by specific *GANs*. They have also shown that augmentations like random *JPEG* compression or Gaussian Noise significantly enhance cross-generator performance. To encourage the CNN classifier to focus on the frequency information, which is a generator-agnostic characteristic, Gragnaniello *et al.* [16] suggest removing the initial downsampling layers of *ResNet50*. Cozzolino *et al.* [17] have combined these concepts but opted for cropping instead of resizing during training. This approach prevents the loss of frequency information and

enhances robustness against resizing augmentations during inference. These detectors are often referred to as universal *GAN* detectors due to their impressive ability to perform well on unseen *GANs* and their robustness against various transformations, making them suitable for real-world scenarios.

Detecting Diffusion Model Generated Images

Unfortunately, several existing studies [6, 7, 9] have indicated that universal *GAN* detectors do not effectively classify images generated by *Diffusion Models* (*DM*). While the frequency spectrum of *DM*-generated images still exhibits differences compared to natural images, these artifacts differ from those present in *GAN*-generated images. Furthermore, these artifacts tend to be more specific to the particular *Diffusion Model* generator, making the achievement of good cross-generator performance a considerably more challenging task. It is also worth noting that *Diffusion Models* are capable of replicating transformation artifacts found in the images they have been trained on with great accuracy [7]. For instance, if the *DM* generator has been exposed to *JPEG*-compressed images during training, the generated images will contain artifacts that are similar to *JPEG*-artifacts. Research also indicates that transformation artifacts in frequency space can closely resemble those generated by *Diffusion Models*. This complicates the task of ensuring robustness to transformations and as our research will demonstrate, it underscores the importance of carefully selecting the training dataset for the detection model.

Nevertheless, it still appears relatively easy to detect images generated by a specific generator, even with small detector models and training datasets [9, 18]. Prior work has shown that *Diffusion Models*, for example, exhibit systematic errors in projective geometry [14]. However, to the best of our knowledge, no detection model has achieved good cross-generator performance and robustness in realistic application scenarios.Promising methods have emerged that achieve outstanding results on small datasets, but they have yet to prove their mettle on larger and more diverse benchmarks like the *GenImage* dataset [9]. *DIRE* [13], for instance, measures the disparity between an image and its reconstructed version of a *Diffusion Model* to detect images from both *Diffusion Models* and *GANs*. They report nearly perfect results in cross-generator performance and robustness on the *DIRE* dataset, but performance drastically declines when evaluated on *GenImage* [10]. *GenDet* [10] significantly improves results by approaching the task as an outlier detection problem instead of binary classification. Effectively, they train a Teacher/Student-network and, to the best of our knowledge, present state-of-the-art results on *GenImage*.

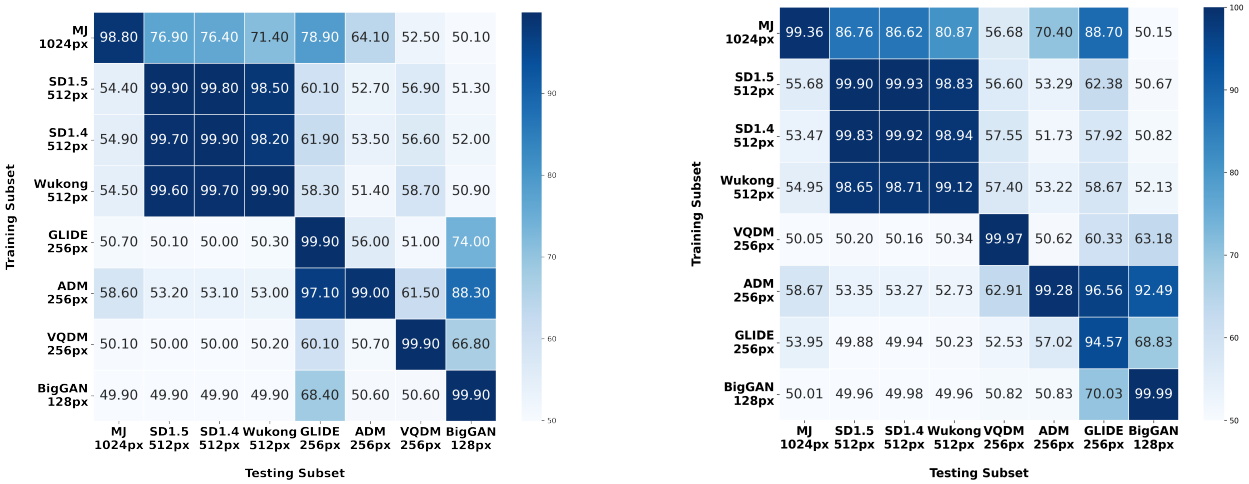
2.2 GenImage Dataset

The *GenImage* dataset [9], stands out as one of the biggest and most diverse datasets for generated image detection. Its primary goal is to establish a unified benchmark within the research community, facilitating the evaluation of detection methods on a standardized dataset. Built upon the *ImageNet* dataset [11], *GenImage* incorporates the original *ImageNet* images as the natural-image class. For the AI-generated class, *GenImage* includes images generated by eight generative models, comprising seven *Diffusion Models* and one *GAN*. These models are specifically *Midjourney* (MJ) [19], *Stable Diffusion V1.5* (SD1.5) [20], *Stable Diffusion V1.4* (SD1.4) [20], *Wukong* [21], *VQDM* [22], *ADM* [23], *GLIDE* [24], and *BigGAN* [25]. They are either Text-to-Image *(MJ, SD, Wukong, VQDM, GLIDE)*, utilizing *ImageNet*-classes as text prompts, or class-conditional *(ADM, BigGAN)*, to guarantee a consistent content distribution between natural and AI-generated images. The dataset is organized into eight distinct subsets, one for every generative model. Each subset contains training and validation data, with a nearly equal number of natural images from *ImageNet* and generated images from the respective generative model. To assess the performance of a detection method, the standard approach involves training on one *GenImage* subset and evaluating on the others. This method enables the measurement of a detection method's cross-generator performance, providing insights into its ability to generalize across different generative models. The evaluation process is quantified by a cross-generator matrix. Fig. 1 illustrates the results presented in the original paper using a basic *ResNet50* classifier, with accuracy (in %) serving as the key metric.

3 Common Biases in Datasets for AI-Generated Image Detection

When examining various datasets used for AI-generated image detection, it is evident that a significant disparity exists in terms of compression techniques. As demonstrated in Table 1, a common practice involves storing all artificially AI-generated images within the dataset as *PNG* files, which involves lossless compression, while natural images are stored in *JPEG* format, involving lossy compression and introducing noticeable artifacts. This prompts us to investigate whether detectors trained on such datasets rely on detecting *JPEG* compression artifacts for classification.

Furthermore, considering that the generators employed for most of AI-generated image detection datasets generate images of a fixed size, as opposed to the diverse size distribution found in natural images, we investigated whether this disparity in size distribution could potentially cause detectors to distinguish between natural and generated images based on their dimensions.



(a) *GenImage* Results (b) Reproduced Results

Figure 1: Reproduced Results. Cross-generator performance of a *ResNet50* classifier from the *GenImage* paper (left), and our reproduced results (right). The matrix shows the accuracy (in %) of a model trained on a *GenImage* training subset (row), when evaluated on a *GenImage* validation subset (column).

The reason why this is problematic is twofold: firstly, it can lead to an overestimation of the performance in detecting generation-specific artifacts, since evaluation data often contains the same biases from which the detector learned during training. Secondly, detectors having such a reliance on undesirable variables do not generalize well to datasets where biases related to these variables are absent, which is the case in real world scenarios. This lack of robustness hinders the adaptability of detectors to changes in these variables.

We conducted experiments using the *GenImage* dataset to assess whether detectors, trained on datasets containing such biases, inadvertently acquire information from these undesirable variables. For these experiments, we trained *ResNet50* detectors since this is the baseline methodology in the *GenImage* paper. We used the provided code from *GenImage* and successfully replicated the reported results, as depicted in Fig. 1. The results show near perfect accuracy when the detectors are tested on the subset containing generated images from the same generative model they were trained on, but poor accuracy on others. By sorting the columns and rows according to the output size of the respective generative models, as we have done, it becomes apparent that better generalization tends to occur near the diagonal.

Table 1: Overview over common AI-Generated Image Detection Datasets and their image compression and size properties.

Dataset	Natural Images			Synthetic Images		
	Compression	Size	Source	Compression	Size	Source
GenImage [9]	*JPEG*	diverse	*ImageNet*	*PNG*	128x128, 256x256, 512x512, 1024x1024	*LAION*, *ImageNet*, unknown
Wang *et al.* [12]	mostly *JPEG* (saved as *PNG*)	diverse (resized to 256x256)	*LSUN*, *ImageNet*, *CelebA*, *COCO*	*PNG*	256x256	trained on same data as natural images
DIRE [13]	*JPEG*	diverse	*LSUN*, *CelebA-HQ*, *ImageNet*	*PNG*	256x256, 512x512, 1024x1024	*LSUN*, *ImageNet*, unknown
Epstein *et al.* [26]	*JPEG*	diverse	*LAION*	*PNG*	256x256, 512x512, 1024x1024	*LSUN*, *ImageNet*, *LAION*, unknown
Ricker *et al.* [6]	*JPEG*	256x256 (resized to)	*LSUN-Bedroom*	*PNG*	256x256	*LSUN-Bedroom*
Ojha *et al.* [8]	*JPEG*	diverse	*ImageNet*, *LAION*	*PNG*	256x256	*ImageNet*, unknown

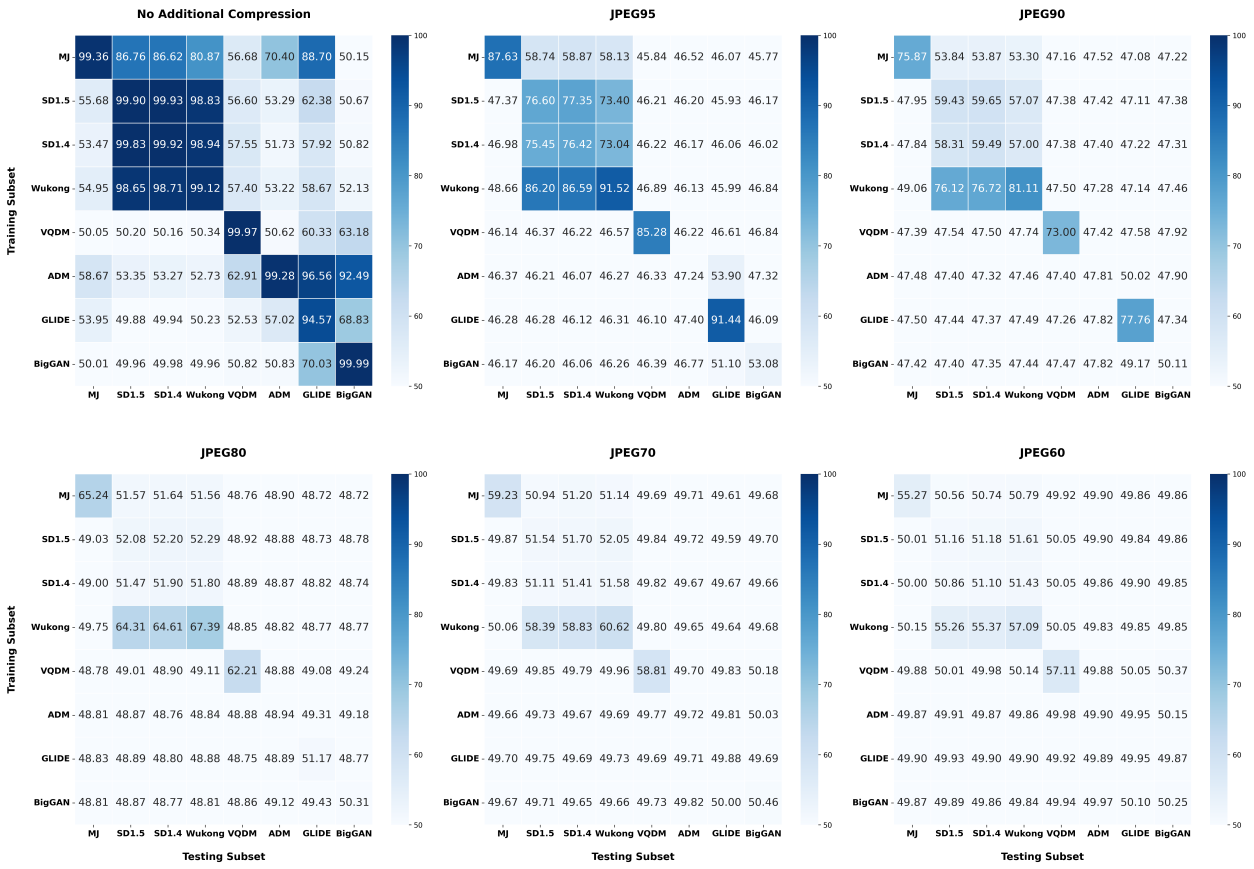


Figure 2: Cross-generator performance of detectors trained on raw *GenImage* for different compression quality factors, given in accuracy (in %).

This observation raises the question of whether similarity in generator size may enhance cross-generator performance, a topic we discuss in section 3.2 and section 4.2.

3.1 JPEG Compression Bias

As mentioned, the natural images used by *GenImage* are sourced from the *ImageNet* dataset. Appendix A illustrates the distribution of *JPEG* quality factors employed in *ImageNet* on a logarithmic scale, with the majority of images compressed using a quality factor of 96 [27]. On the other hand, generated images from *GenImage* are uncompressed. This clear disparity of compression between natural and generated images is common in many datasets, as shown by Table 1.

To investigate whether detectors trained on datasets containing such compression disparities partially function as *JPEG* detectors, we conducted two experiments on the *GenImage* dataset. Initially, we examined whether compressing the dataset's generated images influences their classification as natural. To do so, we used the *ResNet50* detector trained on raw *GenImage* dataset (*cf.* Fig. 1) and evaluated the cross-generator performance of these detectors on test data that was progressively compressed with lower quality factors.

Fig. 2 illustrates a strong decline in accuracy with increased compression, even for high quality factors like 95. Analyzing the confusion-matrix of the results revealed that the precision in detecting AI-generated images consistently remained close to one, whereas the recall dropped significantly (*cf.* App. B). This suggests that compressing a generated image considerably increases the likelihood of the model classifying it as natural.

However, this experiment alone does not confirm that the detector learned to identify *JPEG* compression, as compression might destroy generation-specific artifacts. To address this possibility, we used uncompressed natural PNG images from the FFHQ dataset [28] to observe how compression impacts their classification. Utilizing the same detectors, we evaluated their performance on the FFHQ images as they were *JPEG*-compressed with different quality factors.

Table 2: Accuracy (in %) of the *ResNet50* detector trained on the *Midjourney* subset on 1024×1024 natural FFHQ images when progressively increasing compression.

Compression	*PNG*	*JPEG95*	*JPEG90*	*JPEG80*	*JPEG70*	*JPEG60*
Accuracy	80.45	94.84	98.93	99.95	99.99	100.0

Table 2 shows the results for the detector trained on the *Midjourney* subset of *GenImage*. We observe that the detectors' ability to accurately classify a natural image improves as the level of compression increases. Specifically, the detector's accuracy improves from 80.4% for uncompressed *PNG* images up to 100% for natural images compressed with a low-quality factor of 60. Even a small compression with qualtity factor 95 leads to a much better accuracy of 94.8%. This experience underscores the direct influence of compression towards the classification as natural, as compression, in this scenario, can´t result in the potential destruction of generation-specific artifacts.

3.2 Size Distribution Bias

The generated images in the *GenImage* dataset originate from eight different generators, producing images of four different sizes: 1024×1024 (*MJ*), 512×512 (*SD4*, *SD5*, *Wukong*), 256×256 (*GLIDE*, *ADM*, *VQDM*), and 128×128 (*BigGAN*). In contrast, the natural images sourced from *ImageNet* contain images of various sizes, as shown in Appendix A.

We anticipated that this fundamental difference in size distribution could be exploited by some detectors to discern whether an image is generated or not. For instance, consider a *ResNet50* detector which initially resizes all the input data to 224×224 pixels, as the *ResNet50* detector evaluated in *GenImage*, and is trained with generated images from the *GLIDE* generator. Given that most natural images from the *GenImage* dataset have dimensions around 450×450 pixels, they would typically undergo more resizing than the 256×256 images from *GLIDE*. Consequently, the detector could potentially extract information about the nature of an image by detecting the strength of resizing artifacts. Furthermore, for non-square images, resizing results in significant information loss along one axis, leading to frequency artifacts. However, this problem extends beyond detection methods that rely solely on resizing as a preprocessing step. Detectors employing cropping to achieve a uniform input size are still subject to bias in terms of object size. Discrete-Cosine-Transormation based detectors will contain similar biases as well, as larger images will have more data points (pixels), which affects the frequency spectrum towards having higher frequencies.

To investigate whether detectors trained on the *GenImage* dataset indeed acquire information about image size, we conducted an experiment to evaluate how well a detector performs on natural images of various sizes. A decrease in performance for natural images that closely match the dimensions of the generated images the detector was trained on might suggest that the detector is, to some extent, a size detector. Fig. 3 displays the accuracy of *ResNet50* detectors on natural images across different size intervals. For clarity, we show the performance of detectors trained on one subset for each generator-size available in the *GenImage* dataset. Each diagram illustrates the performance of a detector trained on a specific *GenImage* subset. The analysis included all natural *ImageNet* images not seen by the detector during training, effectively using the training as well as validation data from other *GenImage* subsets. Appendix A illustrates the number of *ImageNet* images in each interval. It is important to highlight that some intervals contain very few images, which is why the single cells are not as informative as the global trend. Note that we used the models trained in section 4.1, since otherwise the classification of natural images is highly affected by *JPEG*-artifacts and thus the prediction is near perfect for all intervals.

Our findings indicate a correlation between decreased detector performance and natural images sized similarly to those of the generated images the detector was trained on. These results demonstrate that detectors indeed perform better on natural images that have significantly different sizes compared to the size of the generated images used during detector training. For instance, the *BigGAN* detector accurately classifies most natural images, except from very small images with one side between 100-150, as *BigGAN* images are of size 128×128. *ADM* images are of size 256×256, which is why the effect is also visible for slightly bigger images. Conversely, the *Midjourney* detector shows reduced performance on larger natural images, given that *Midjourney* images are much bigger than most natural images in *ImageNet*. For *Stable Diffusion*, no significant trend emerges, as the generated images are of similar size as most of the natural images. This detector has to learn more from other discriminative patterns in the interval of the generated images, which does not mean that the the detector did not learn from the bias to classify images outside of the interval.

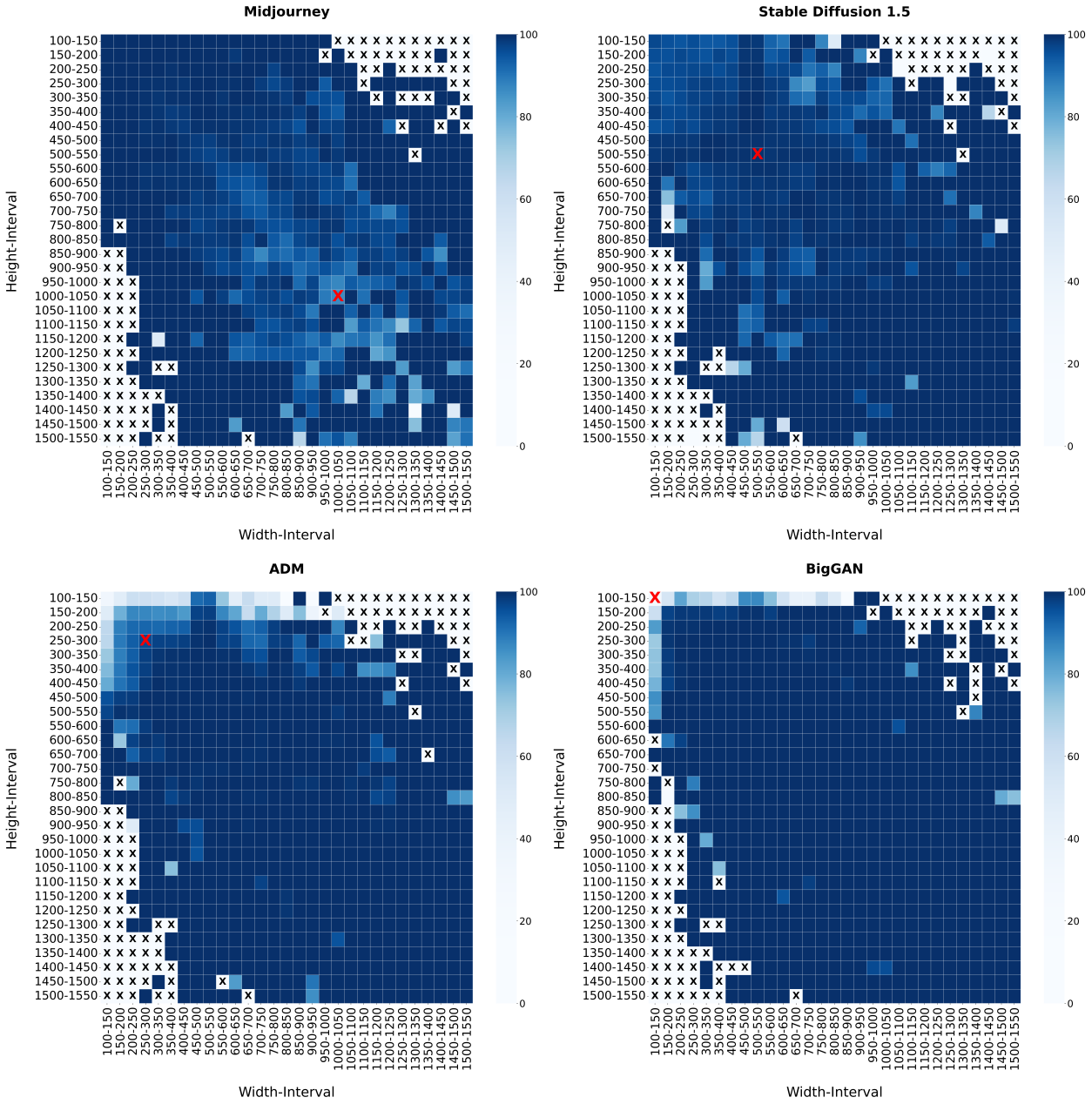


Figure 3: Accuracy (in %) of detectors trained on different *GenImage* subsets when evaluated on natural images of varying sizes. Black crosses indicate intervals without any data. The red cross marks the size interval of generated images corresponding to this subset.

4 Removing The Biases

We have emphasized that disparities in size distribution and compression between natural and generated data can cause detectors to learn from differences that may not be pertinent in real-world contexts. Consequently, we retrained the same detectors using a constrained dataset and reevaluated them to investigate whether the removal of compression and size biases would affect the final evaluation of detectors. Our findings reveal significantly divergent results and notably enhanced cross-generator performance and robustness.

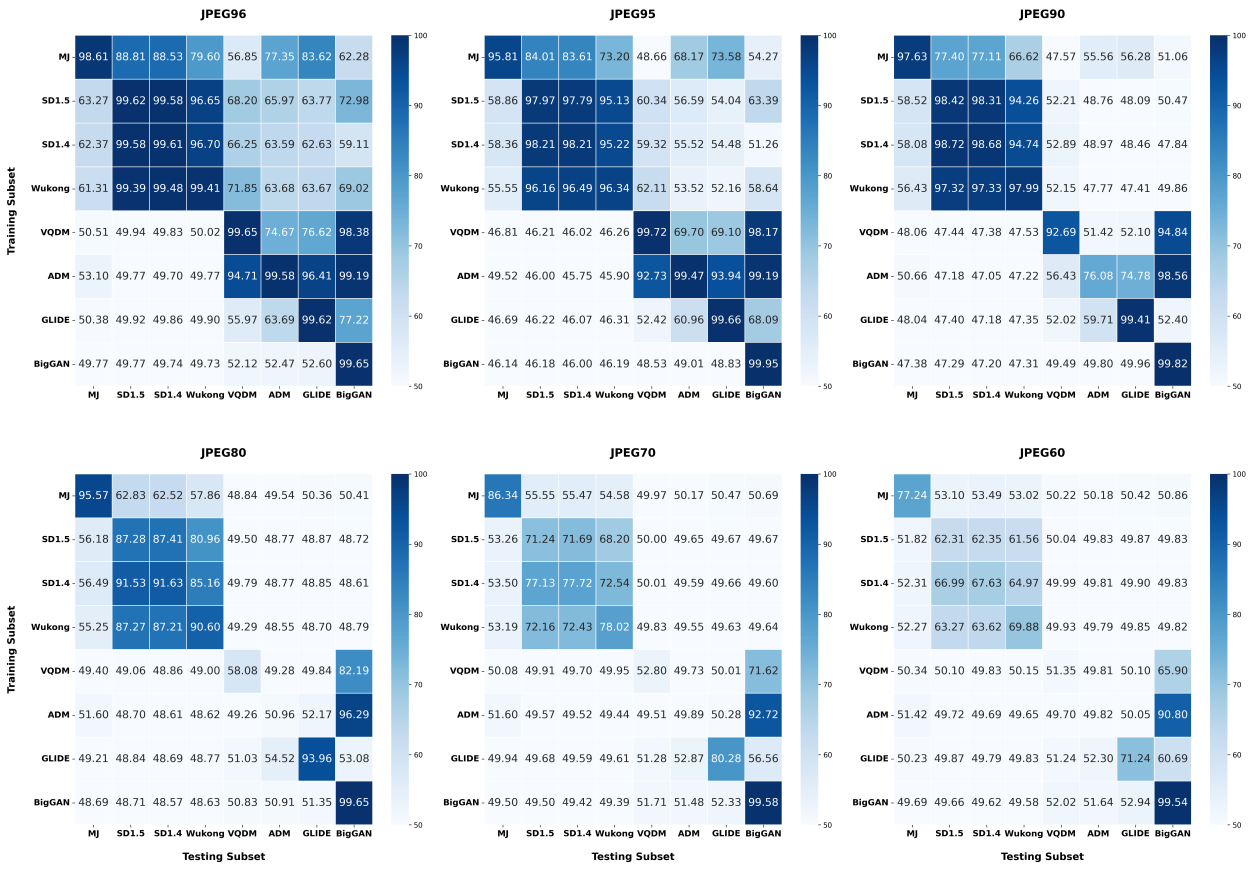


Figure 4: Cross-generator performance given in accuracy (in %) of detectors trained only with images compressed with *JPEG* quality factor of 96 from *GenImage* for different Compression Rates.

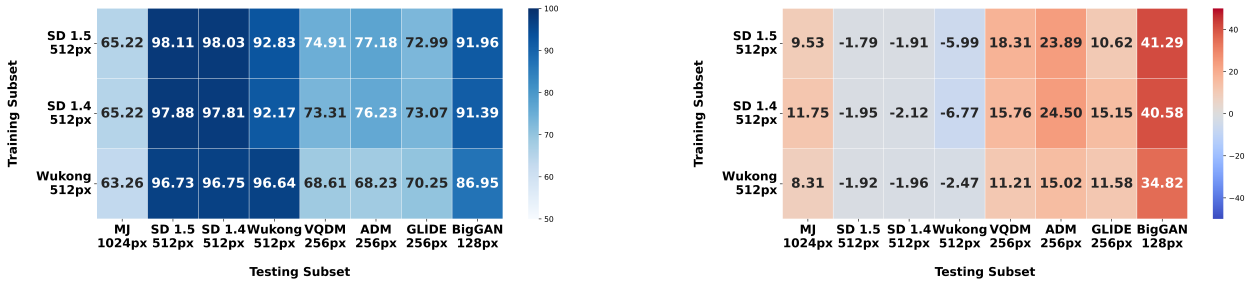
Table 3: Comparison of robustness to compression between training on raw *GenImage* and only *JPEG96 GenImage*: Accuracy in % averaged cross the generator matrix.

	Training dataset		
Compression	**Classic *GenImage***	***JPEG96* (ours)**	**Difference**
JPEG95	53.91	67.17	+13.26
JPEG80	50.62	59.37	+8.75
JPEG60	50.58	55.07	+4.49

4.1 JPEG Constraint

To ensure that the detector does not learn to differentiate between natural and generated data based on *JPEG* compression artifacts, we constrained the quality factor of all images used in training. Specifically, we constructed a training set by exclusively selecting natural images compressed with a quality factor of 96. We then selected an equivalent number of generated images and compressed them using the same quality factor. Subsequently, we trained a *ResNet50* detector using this constrained dataset and reevaluated it's cross-generator and robustness performance. This approach is expected to make the detector's classification less sensitive to *JPEG* compression than before. Fig. 4 illustrates the cross-generator performance of these detectors when the test data is compressed with increasing quality factors. Unlike the training data, the test data is not subject to any constraints. It is important to note that for *JPEG96*[1], compression was applied solely to the generated images, as the majority of natural images is already compressed to *JPEG96* or a lower quality. Consequently, for compression rates other than *JPEG96*, natural images undergo a second round of compression.

[1]We mean with JPEGX a JPEG compression with a quality factor of X.



(a) Accuracy (in %) for size-constrained training on full validation datasets

(b) Difference to training on raw *GenImage* dataset

Figure 5: Cross-generator perfomance for *ResNet50* detectors trained on compression and size constrained *GenImage* training subset.

Comparing the results shown in Fig. 4 with those in Fig. 2, which presents the same evaluation for detectors trained on the raw *GenImage* dataset, we observe a strong enhancement in robustness against *JPEG* compression. Specifically, as illustrated in Table 3, we observe an overall improvement of 13.26 accuracy points to *JPEG95* compression, 8.75 for *JPEG80*, and 4.49 for *JPEG60*. For example, the *BigGAN* detectors performance remains nearly intact up to *JPEG60*, whereas previously, the detector incorrectly classified all *JPEG95* images as natural.

These results prove our assumption that biases in compression lead to the detectors learning wrong causalities and not being robust to changes in compression. Previous research [12] has already indicated that introducing random *JPEG* augmentation into the training set enhances detector robustness and generalization to other generators. We firmly believe that this improvement does not solely stem from increasing the training set's variance but also partially from reducing the bias arising from differences in *JPEG* compression during training. Nevertheless, random-augmentation does not completely remove the bias, since it merely shifts the distributions of quality factors to have a bigger overlap instead of equalizing them. Moreover, when using augmentation, already compressed images are compressed a second time, while uncompressed images are only compressed once. This introduces a bias, as *JPEG* compression is not an idempotent transformation.

Surprisingly, mitigating the *JPEG* bias marginally improves the generalization across different generators. This contradicts our initial expectation that the bias, being present in all *GenImage* subsets, should facilitate transferability. To explain this we propose two hypothesis: Firstly, a training dataset with reduced bias compels the detectors to focus on learning from generation-specific artifacts. These artifacts are not only more resilient to compression but also to other transformations like resizing, thereby enhancing generalization to generators of various sizes. Secondly, research indicates that *Diffusion Models*, when trained on *JPEG*-compressed images, produce artifacts that mimic those of compression artifacts [7]. This suggests that biased detectors do not actually distinguish between compressed and uncompressed images, but rather between genuine compression artifacts and their approximations. This distinction may generalize less effectively to other generative models.

4.2 Size Constraint

To minimize the detector's reliance on variations in size, we aimed to control this variable by selecting natural images within a specific size range for training, such that the natural images in training are of similar size as those of the corresponding generator. As this selection drastically reduces the number of training samples, we only executed the experiment on subsets with generated images of dimension 512×512, because most natural *GenImage* images have both height and width within the range [450, 550]. In order to obtain sufficient training data, we also utilized the natural training data from all *GenImage* training subsets in this interval. Furthermore we still only used *JPEG96* images to mitigate the compression bias. We then sampled the same number of generated images for each 512×512 generator. To avoid disparities in content distribution between natural and generated images, we ensured an equal number of natural and generated images per *ImageNet* class. This data selection in total reduced the number of training samples from approximately 320,000 to 75,000. For the evaluation, we utilized the unconstrained validation sets, to maintain consistency with other experiments. To guarantee that all images images contain the same amount of resize artifacts during training, we center-cropped images to 450 - the lower bound of the size interval - before resizing to 224 the final inputsize of the *ResNet50* detectors. During inference, for images of any size, we first resize the input to 512 to allow cropping to 450 and then resizing.

Table 4: Average cross generator performance for *ResNet50* and *Swin-T* given in accuracy (in %) when trained on raw *GenImage* subsets and our constrained subsets.

Training Subset	ResNet50			Swin-T		
	Classic	Ours	Diff	Classic	Ours	Diff
SD1.5	72.16	83.90	+11.74	74.14	85.90	+11.76
SD1.4	71.27	83.39	+12.12	74.93	86.80	+11.87
Wukong	71.61	80.93	+9.32	73.20	84.80	+11.60
Total	71.68	82.74	+11.06	74.09	85.83	+11.74

Fig. 5a shows the cross-generator performance of detectors trained with this size constraint, while Fig. 5b highlights the difference compared to detectors trained on the unconstrained, raw dataset. Note that for the evaluation, images stored as *PNG* files are compressed using a *JPEG* quality factor of 96, ensuring the model is not exposed to uncompressed images, which it had not encountered during training. The results demonstrate an overall improvement in cross-generator generalization, with a maximum increase of 41.29 percentage points and an average increase of 11.06 percentage points, achieving state-of-the-art results in *GenImage* cross-generator performance. We note a slight decrease in performance for subsets containing generated images of the same size as those used in training. This observation supports our hypothesis, since the original evaluation likely overestimated the generalization ability for generators of the same size: For instance, detectors trained on the stable-diffusion subset performed exceptionally well on the *Wukong* subset, which shares the same size bias. Mitigating the size bias leads to a decline in generalization from stable-diffusion to *Wukong* of approximately 7 percentage points.

4.3 Detector Ablation

To ensure that the constraints not only improve the performance of *ResNet50* detectors, we likewise examined our constrained training for a transformer based detector *Swin-T* [29]. Table 4 summarizes the average scores: Note that we outperform the current stat-of-the-art, *GenDet* [10], which only reports results of 81.6% for training on *SD1.4*. When validating our *Swin-T* detectors, we achieve an average accuracy of 85.83%. Specifically, the detector trained on the *SD1.4* subset reaches an accuracy of 86.8%. We provide the detailed cross-generator matrix of *Swin-T* in Appendix C. The overall improvement in cross-generator performance is especially impressing when considering that the selection of training data reduced the number of samples by more than 75%. Nevertheless, the key insight of our work is not merely enhanced performance and robustness but rather showing that the named biases significantly affect the detectors leading to a misjudgment when evaluating.

5 Discussion and Conclusion

Our findings demonstrate that datasets used for AI-generated image detection exhibit biases in compression artifacts and image dimensions. These biases hinder models from learning the core task of detecting generation specific artifacts and result in misaligned evaluations. By imposing constraints on the training dataset to mitigate these biases, we observed a significant shift in the evaluation of *ResNet50* and *Swin-T* detectors, yielding substantially improved robustness and generalization. Since we only inspected these baseline methods, it would be interesting for future works to analyze if we can improve the performance of detection methods specifically designed for AI-generated image detection even more or whether they expose as only being able to perform well on biased datasets. While computational and time limitations prevented our evaluation of the *DIRE* method yet, it remains a promising candidate for further investigation.

Nevertheless, it's important to highlight that even with these constraints, datasets may contain other undesirable biases. A notable issue is the disparity in the source of natural images used for training generative models compared to those used for detector models, as shown in Table 1. For instance, in the *GenImage* dataset, *Stable Diffusion* generators are trained on images from *LAION*, whereas detector models use images from *ImageNet*. Despite efforts to align image content through text prompts corresponding to *ImageNet* classes, there is a potential risk that detectors might learn to distinguish between the styles of *LAION* and *ImageNet* images. Even more concerning, some generative models like *Stable Diffusion* actively fingerprint the generated content [30], leading to a undesirable distinction from the natural images of the detector that is clearly not transferable to other generative models.

To create truly unbiased datasets, we propose that the detector models should be trained on the same natural images as their corresponding generative model. Additionally, natural and generated images should have near equal distributions in both compression and image dimensions.

References

[1] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in Neural Information Processing Systems*, 36, 2024.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.

[4] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[6] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

[7] 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*, pages 973–982, 2023.

[8] 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.

[9] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang. Gendet: Towards good generalizations for ai-generated image detection. *arXiv preprint arXiv:2312.08880*, 2023.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] 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.

[13] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023.

[14] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. *arXiv preprint arXiv:2311.17138*, 2023.

[15] 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.

[16] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2021.

[17] Davide Cozzolino, Diego Gragnaniello, Giovanni Poggi, and Luisa Verdoliva. Towards universal gan image detection. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021.

[18] Sergey Sinitsa and Ohad Fried. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4067–4076, 2024.

[19] Midjourney. `https://www.midjourney.com`, 2022.

[20] Stable Diffusion WebUI. `https://github.com/AUTOMATIC1111/stable-diffusion-webui`, 2022.

[21] Wukong. `https://xihe.mindspore.cn/modelzoo/wukong`, 2022.

[22] 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.

[23] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[24] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021.

[25] A Brock, J Donahue, and K Simonyan. Large scale gan training for high fidelity natural image synthesis. international conference on learning representations. 2019.

[26] David C. Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *ICCV DeepFake Analysis and Detection Workshop*, 2023.

[27] towardsdatascience. Compression in the ImageNet dataset. `https://towardsdatascience.com/compression-in-the-imagenet-dataset-34c56d14d463`.

[28] 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*, pages 4401–4410, 2019.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[30] Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang. Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. *arXiv preprint arXiv:2306.04744*, 2023.

Appendix

Appendix A shows detailed insights into size (Fig. 7)and compression (Fig. 6) distribution in *ImageNet* images, which are used in *GenImage*. Appendix B shows the precision and recall corresponding to the experiment in section 3.1 to show that compression leads to classifying an image as natural. Appendix C provides the detailed cross-generator matrix for the *Swin-T* detector when trained on the constrained dataset and it's difference to training on the raw *GenImage* dataset.

Appendix A

Figure 6: Distribution of *JPEG* quality factor in *ImageNet* with logarithmic scale. (Graphic from [27])

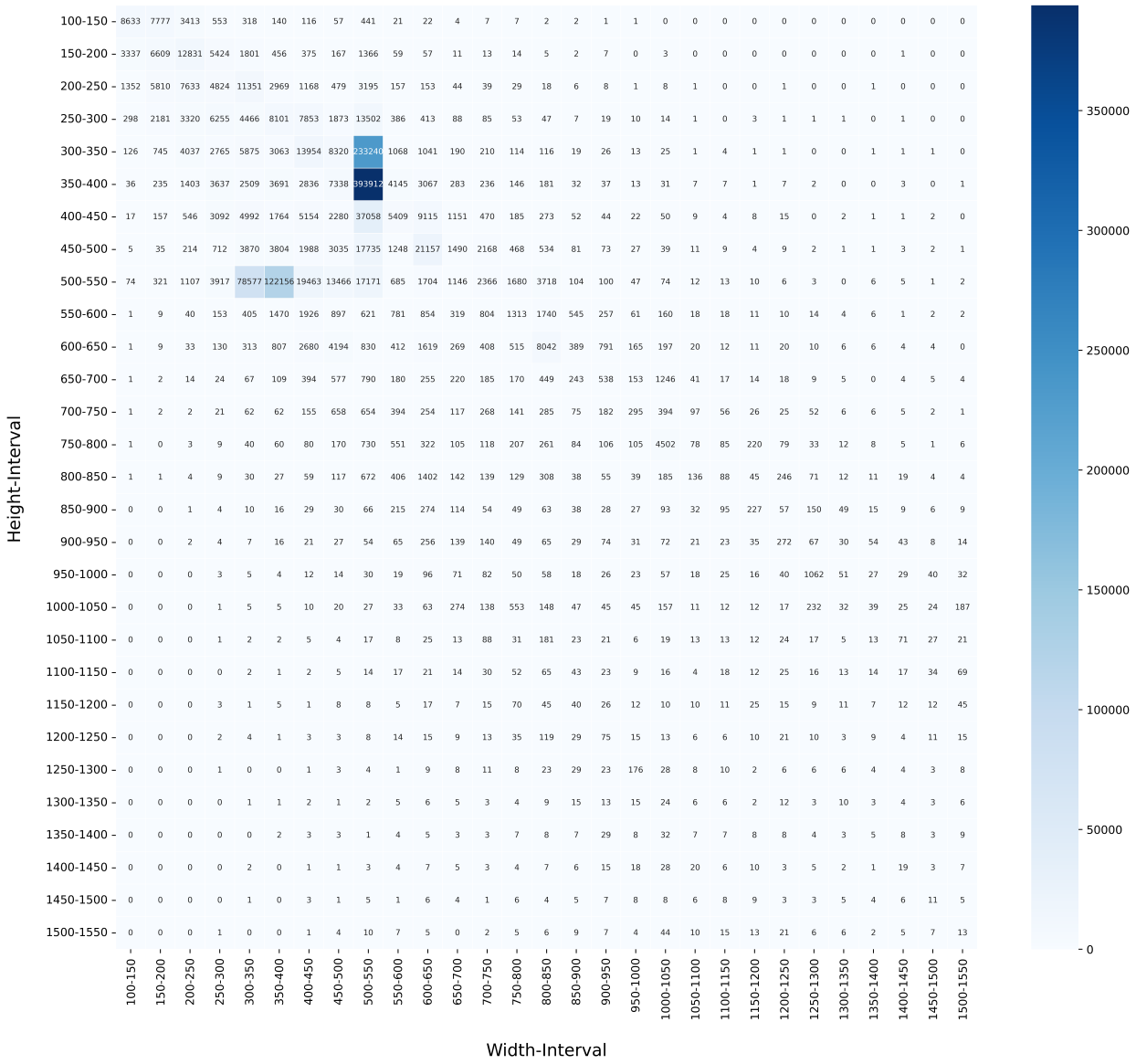


Figure 7: Distribution of image size in *ImageNet* per interval.

Appendix B

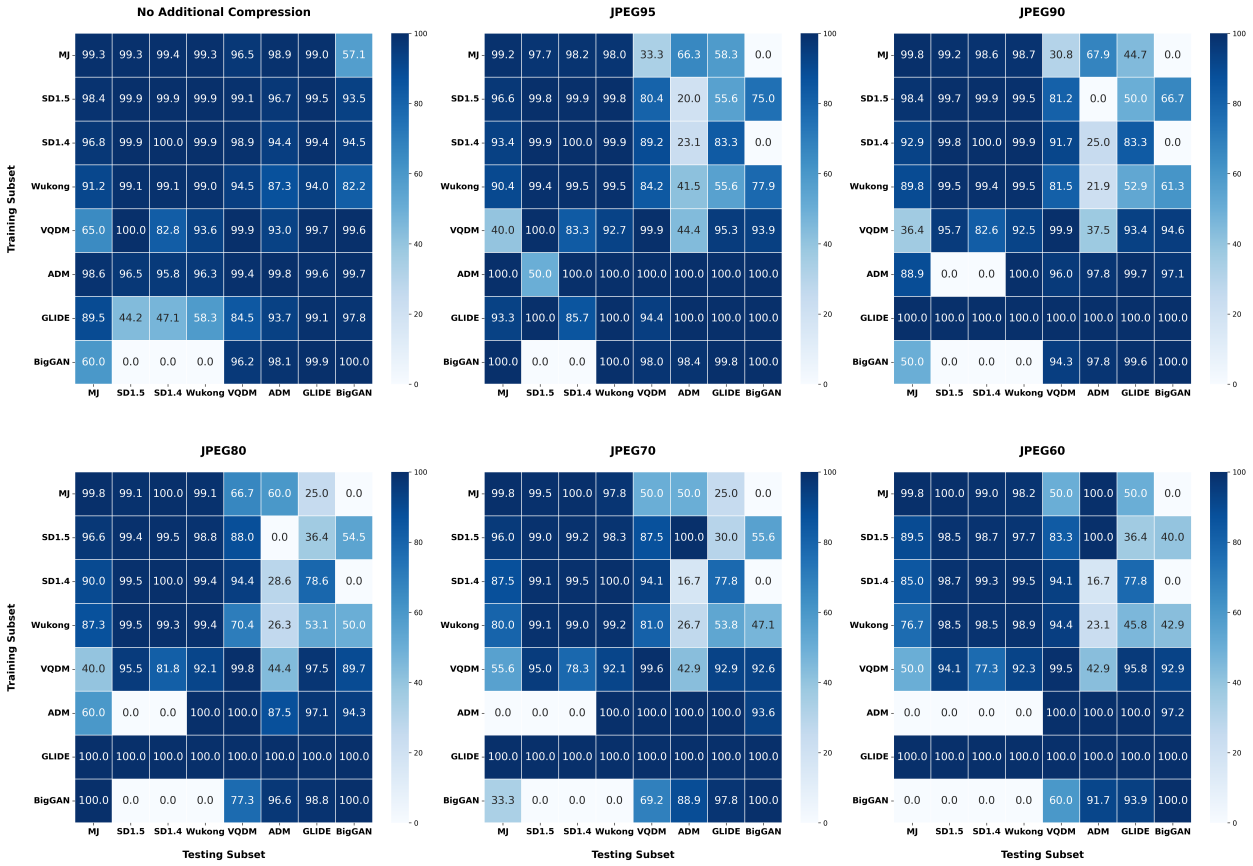


Figure 8: Precision (in %) corresponding to Fig. 2. This demonstrates that the classification of natural images is not affected by the compression.

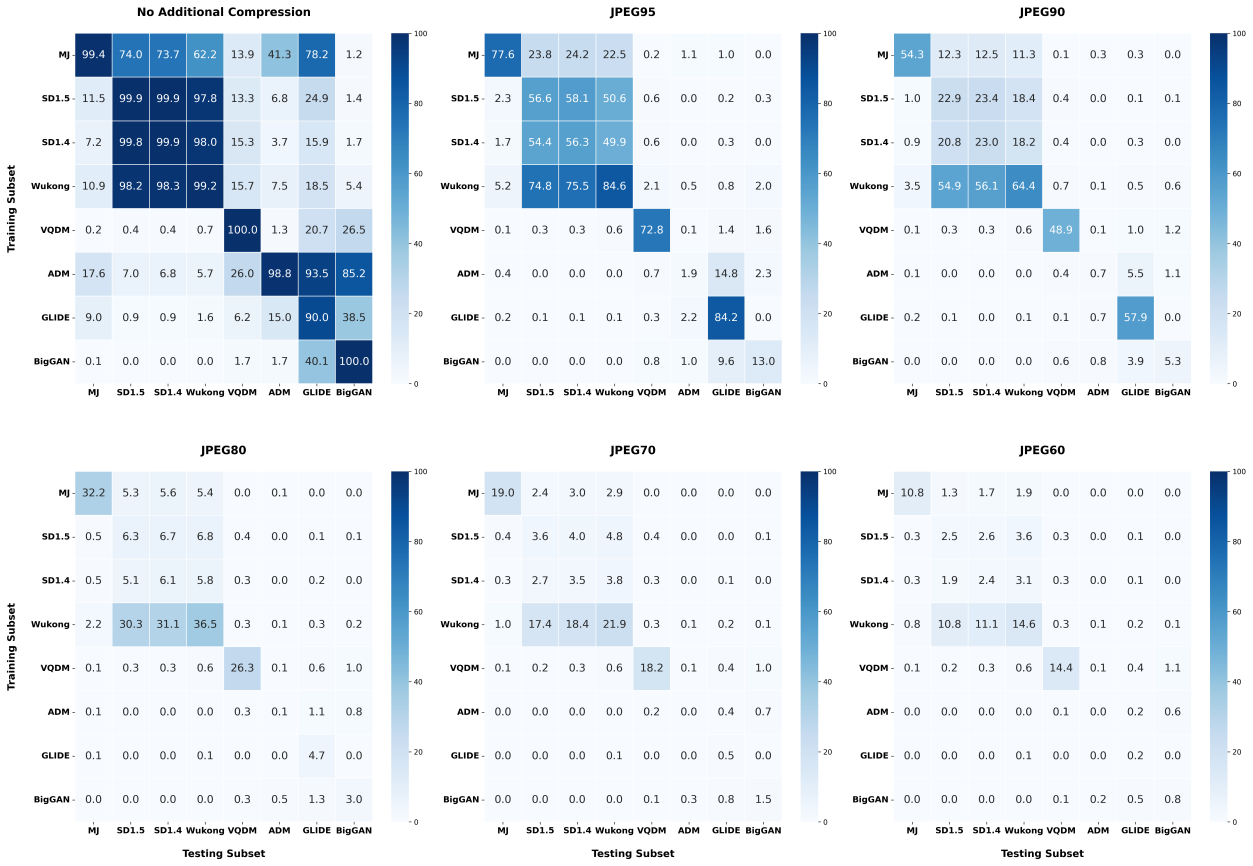
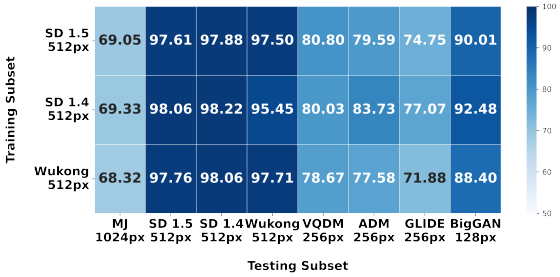
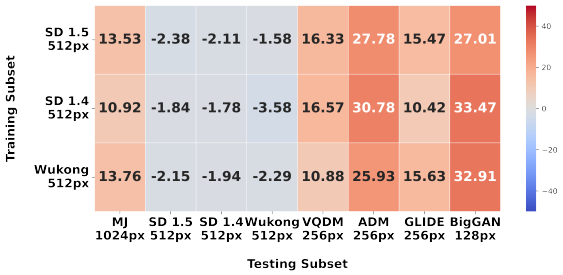


Figure 9: Recall (in %) corresponding to Fig. 2. This demonstrates that compressed AI-generated images are likely classified as natural

Appendix C



(a) Accuracy (in %) for constrained training in 4.2. on full validation datasets



(b) Difference to training on raw *GenImage* dataset

Figure 10: Cross-generator perfomance for *SWIN-T* detectors trained on compression and size constrained *GenImage* training subset.