Present and Future Generalization of Synthetic Image Detectors

Pablo Bernabeu-Pérez (✉), Enrique Lopez-Cuena, and Dario Garcia-Gasulla

Barcelona Supercomputing Center (BSC)
pablo.bernabeu@estudiantat.upc.edu, {enrique.lopez,dario.garcia}@bsc.es

Abstract. The continued release of increasingly realistic image generation models creates a demand for synthetic image detectors. To build effective detectors, we must first understand how factors like data source diversity, training methodologies and image alterations affect their generalization capabilities. This work conducts a systematic analysis and uses its insights to develop practical guidelines for training robust synthetic image detectors. Model generalization capabilities are evaluated across different setups (*e.g.,* scale, sources, transformations), including real-world deployment conditions. Through extensive benchmarking of state-of-the-art detectors across diverse and recent datasets, we show that while current approaches excel in specific scenarios, no single detector achieves universal effectiveness. Critical flaws are identified in detectors and workarounds are proposed to enable practical applications that enhance accuracy, reliability and robustness beyond current limitations.

Keywords: synthetic image detection · ai-generated images · diffusion models · model generalization · detector robustness.

1 Introduction

Synthetic image generation presents challenges regarding visual information integrity, misinformation mitigation and trust in digital environments. Due to these concerns, correctly attributing synthetic content has become a social demand and a top scientific priority. Recent legislation aligns with this context, mandating the identification and notification of synthetic digital content [39].

To address these needs, synthetic image *detection* (SID) has become locked in a race with synthetic image *generation* (SIG) [25]. SID aspires to win by developing universal detectors [32, 8], but their generalization capacity remains uncertain. Meanwhile, new SIG models join the race every month, advancing in realism and posing new challenges. This work studies the SIG-SID relationship by analyzing the impact of training conditions on SID generalization (§4). The lessons learned are applied to train a baseline for evaluating generalization under deployment conditions, including variations in data and model sources and scaling factors (§5). Our final experiments (§6) benchmark recent detectors using synthetic data from the latest generators under optimized image scaling policies.

Finally, ethical considerations related to SID research, including when and how detectors should be publicly released, are discussed (§7).

Our findings indicate that current methods are insufficient for reliable SID, as no tested model generalizes universally. Factors like rescaling play a major role in detector performance, exposing a vector of attack for malicious actors. While some models suffer major degradations, others benefit from resized inputs, emphasizing the importance of appropriate preprocessing techniques. Lastly, detectors perform much worse on private models, like `DALLE` and `Midjourney`, compared to open models, highlighting the crucial role of open science in synthetic attribution. This work illustrates how, as of today, generalization should never be assumed in the field of SID. In summary, our contributions include:

- A systematic analysis of training conditions affecting detector generalization, showing improved robustness when trained on newer generators and highlighting vulnerabilities to common image alterations.
- Development and release of `SuSy`, a multi-class detector trained with optimized augmentations, accompanied by practical deployment guidelines including optimal patch aggregation and standardized rescaling.
- Comprehensive benchmarking of state-of-the-art detectors across diverse datasets, identifying critical vulnerabilities related to image rescaling.

2 Related Work

Previous work on SID has primarily focused on GAN-generated content [48, 40, 18], due to its historical prevalence and relative speed. However, recent studies show that GAN-focused detectors often fail to identify content from modern diffusion models [41, 30]. While several recent works address the detection of diffusion-based content [3, 11, 46, 28, 51, 50, 19, 32], which now produce the most perceptually convincing synthetic images, their generalization ability under diverse conditions remains largely untested.

Deep learning architectures such as CNNs [9, 37] and Visual Transformers (ViTs) [3, 28] have been used to learn hierarchical synthetic patterns, with CLIP-based methods enhancing detection through semantic [32] and intermediate feature analysis [23]. Models combining textual and visual features have also been adopted; [7] uses prompt tuning to detect deepfakes as a visual question-answering problem, while [44] applies contrastive learning guided by text. Hybrid models combine multiple detection signals, such as dual-stream networks analyzing texture and frequency artifacts [45] or CLIP features fused with low-level image statistics [35]. Across detection methods, frequency domain-based approaches are commonly used to reveal generation artifacts [10]. Some leverage Fast Fourier Transform analysis to capture characteristic patterns [36, 5], while others explore wavelet-based features specifically for diffusion outputs [14]. In addition, local feature analysis examines texture contrast patterns [49] and intrinsic dimensionality properties [30] for complementary signals.

AI-generated image detectors are typically trained using data from a single source and evaluated across multiple sources to assess generalization [11, 33, 51,

50, 19, 32, 6, 8]. Among various bias sources examined, image format and resolution stand out as key factors. In [11], authors highlight the impact of resizing operations, a common practice to adjust images to network input resolution. The study in [19] demonstrates biases associated with *JPEG* compression and image size, with detectors generally performing better on natural images that differ significantly from generated training images. This observation aligns with findings in [12], where dataset choice significantly impacts detection performance.

Recent efforts, such as the SIDBench framework [34], have performed SID evaluation across diverse datasets, including an analysis of resolution effects. Part of our work builds on the SIDBench framework but differs in scope and contribution. Firstly, we provide an analysis of resolution effects by examining a range of scaling factors, while SIDBench focuses on cropping versus resizing. We also investigate additional generalization factors, including generator family, model release date and dataset source, including both open and private models, as well as multiple sources for the same generators. Furthermore, we incorporate more authentic datasets and newer synthetic image generators. Lastly, we evaluate optimal scaling settings for individual detectors and benchmark their performance under real-world conditions, uncovering new insights.

3 Methods

To examine detector biases arising from training methodology, we employ a fixed architecture (see §3.1), train it using six image datasets (see §3.2) and evaluate it with fifteen additional datasets (see §3.3). To enable full reproducibility of our work, our codebase [1], training datasets [2] and model weights [3] for our best-performing detector are publicly released.

3.1 Architecture

For our experimentation, we use a ResNet [22] trained as a direct classifier, chosen for its robust performance [51, 19, 6] and lightweight design suitable for large-scale evaluation. Specifically, we adopt the staircase design from [31], which combines CNN-based feature extraction with MLP classification in a staircase design shown in figure 1 (see Appendix A for a detailed explanation of the architecture).

Detectors are commonly trained on image patches or downsampled images, as processing entire high-resolution images is computationally intensive and the most discriminating features are typically low-level. For each image, we select the five 224×224 patches exhibiting the highest contrast in their grey-level co-occurrence matrix [21]. These patches are processed individually through the

[1] https://github.com/HPAI-BSC/SuSy
[2] https://huggingface.co/datasets/HPAI-BSC/SuSy-Dataset
[3] https://huggingface.co/HPAI-BSC/SuSy

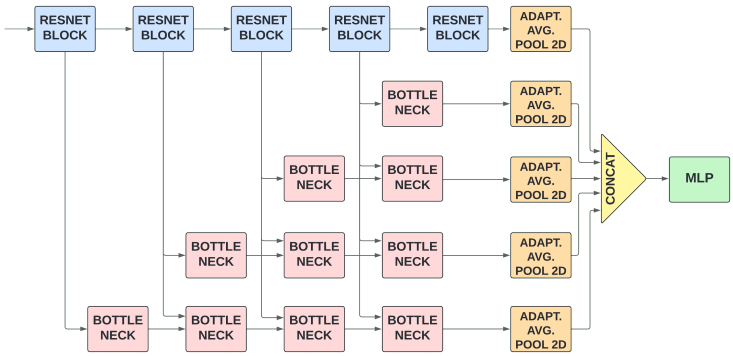

Fig. 1: Detector architecture used, based on a ResNet-18 from [31], including ResNet blocks (blue), bottlenecks (red), adaptative average pooling 2D (orange), concatenation (yellow) and an MLP (green).

network, producing per-patch predictions that are aggregated to yield image-level decisions. The impact of different aggregation strategies on detection performance is analyzed in §5.2.

For performance metrics, recall is used for single-dataset evaluations (authentic or synthetic), focusing solely on the model's ability to identify the target class. For multi-dataset classification scenarios, macro accuracy is employed to provide an unweighted mean of per-class accuracy, ensuring fair evaluation across all classes regardless of sample size.

3.2 Train Datasets

The training experiments detailed in §4 utilize two types of datasets: *authentic* real-world images sourced from `COCO` [27] and *synthetic* AI-generated images from `DALLE3` [16], `SD1.X` [42], `SDXL` [15], `MJ 1/2` [38] and `MJ 5/6` [17]. These datasets represent different versions of three popular image generators: `DALLE`, `StableDiffusion` and `Midjourney`. To ensure balanced class representation, `COCO` and `SD1.X` are undersampled to a maximum of 5,435 images. Pre-existing train, validation and test splits are respected, defaulting to a standard 60%-20%-20% random split when such partitions are unavailable. For `SDXL`, the *realistic-2.2* split is used for training and validation, while the *realistic-1* split is reserved for testing. Further details are provided in Appendix C.

3.3 Benchmarking Datasets

To evaluate SID models, we use fifteen datasets: eleven produced and gathered by others, two produced by others but gathered by us and two produced by us. Image resolution distributions and visual samples are detailed in Appendices G and H, respectively.

Dataset	Model	Year	Format	Type	Train	Val	Test
COCO	-	2017	JPG	Auth	2,967	1,234	1,234
dalle3-images	DALLE3	2023	JPG	Synth	987	330	330
diffusiondb	SD1.X	2022	PNG	Synth	2,967	1,234	1,234
SDXL	realisticSDXL	2023	PNG	Synth	2,967	1,234	1,234
mj-tti	MJ 1/2	2022	PNG	Synth	2,718	906	906
mj-images	MJ 5/6	2023	JPG	Synth	1,845	617	617
Evaluation Datasets							
Flickr30k	-	2014	JPEG	Auth	-	-	31,655
GLDv2	-	2020	JPEG	Auth	-	-	5,000
In-the-wild	-	2024	Mix	Auth	-	-	121
Synthbuster	Multiple	2024	PNG	Synth	-	-	9,000
SD3	SD 3	2024	PNG	Synth	-	-	8,192
FLUX.1	FLUX.1	2024	PNG	Synth	-	-	8,192
In-the-wild	?	2024	PNG	Synth	-	-	99

Table 1: Datasets with generative models, release date, image format, type and sample counts.

The externally produced datasets include two subsets of 5,000 randomly selected authentic images: scenes depicting people from Flickr30k [47] and natural and human-made landmarks from GLDv2 [43]. Additionally, nine synthetic datasets from the Synthbuster superset [5] provide 1,000 images each, generated using common prompts across both models included in our training (*e.g.,* SDXL, DALLE3) and models outside our training set (*e.g.,* DALLE2, Firefly).

The In-the-wild dataset contains images gathered from online sources by the authors. The authentic split includes 121 manually curated images from sources like *Reddit* communities prohibiting AI content and *Flickr* uploads prior to 2020. The synthetic split consists of 99 photorealistic AI-generated images sourced from *Civitai* and *Reddit*'s synthetic content communities. Despite careful manual curation and community moderation, we acknowledge a residual risk of contamination due to possible mislabeling or oversight.

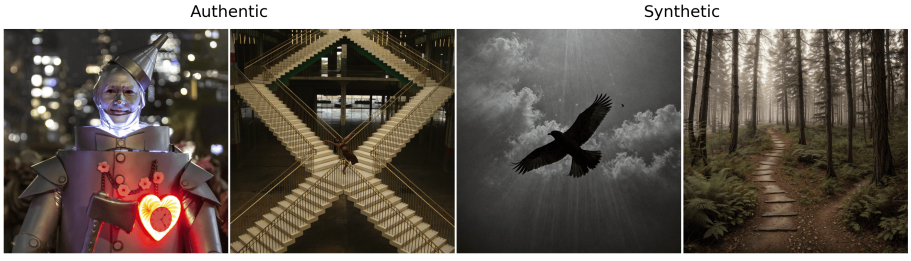

Fig. 2: Examples of the In-the-wild dataset.

Finally, we generate two synthetic datasets containing 8,192 images each: `SD3`, created using Stable Diffusion 3-Medium [4], an MMDiT text-to-image model; and `FLUX.1`, generated with `FLUX.1-dev` [24], a 12B parameter model combining MMDiT and DiT architectures [26]. Additional details are in Appendix D.

4 Train Experiments

This section examines how different training strategies affect model generalization. For consistent experimental comparison, all models share identical architecture (§3.1) and hardware setup (Appendix B). Concretely, all experiments were conducted on the MareNostrum 5 supercomputer, hosted at the Barcelona Supercomputing Center (BSC). We utilize an Intel Xeon Platinum 8460Y processor and one NVIDIA Hopper H100 64GB GPU. Training is capped at 20 epochs with a 2-epoch patience early stopping based on validation accuracy. The datasets described in §3.2 are augmented using horizontal flips with 50% probability, while additional transformations are analyzed in § 4.3.

4.1 Single-class Models

We evaluate relationships between SIG models by training binary classifiers, using each synthetic dataset in §3.2 as a positive class and `COCO` as the negative class. These single-class detectors are then tested on the remaining datasets to assess cross-model generalization (see Table 2).

Training Dataset	**Year**	**Evaluation Dataset**					**Avg.**
		`DALLE3`	`SD1.X`	`SDXL`	`MJ 1/2`	`MJ 5/6`	
`DALLE3`	2023	**97.70**	27.59	70.19	50.73	97.02	68.64
`SD1.X`	2022	49.76	**98.30**	68.23	39.65	40.36	59.26
`SDXL`	2023	51.27	33.45	**97.57**	59.14	67.97	61.88
`MJ 1/2`	2022	31.39	17.63	73.14	**99.07**	51.51	54.55
`MJ 5/6`	2023	91.76	26.13	64.75	62.69	**99.25**	68.92
Avg.		64.38	40.62	74.78	62.26	71.22	

Table 2: Patch-level recall (%) of single-class models for synthetic datasets. In bold, performance on the training dataset.

While single-class detectors achieve excellent recall (over 97%) on their target class, performance drops substantially when tested on other datasets. SIG model age emerges as the dominant factor affecting generalization. When evaluating newer detectors on older generators, we observe severe performance degradation, as shown in the last row of Table 2, where detectors trained on `SD1.X` and `MJ 1/2` (both from 2022) show the lowest average values. This pattern likely stems from older generators producing more pronounced artifacts, which newer detectors struggle to identify without specific training. Conversely, detectors trained on

recent datasets show better cross-SIG generalization, as evidenced by the higher average values for DALLE3, SDXL and MJ 5/6 in the last column. Paradoxically, this suggests that more realistic generators enhance the robustness and reduce the bias of detectors. In contrast, SIG family has a weak effect on generalization - the detector trained on SDXL is below average when tested on SD1.X and the SID model trained on MJ 1/2 is not particularly accurate on MJ 5/6. The effect of image format is also inconclusive.

4.2 Multi-class Models

Multi-class detectors offer richer decision boundaries compared to single-class detectors, which tend to collapse [13] *i.e.,* defaulting to predicting only one class. To explore the effects of this distinction on generalization, we train a binary classifier merging all synthetic data sources from §3.2 into a single synthetic class, including 14,323 synthetic images and an analogous amount drawn from COCO to compose the authentic class. We also train a six-class recognition model using the original splits defined in §3.2. To obtain binary classifications from the six-class model, we take *argmax* of the output probabilities, where all samples labeled as belonging to a synthetic class are considered equal predictions of the synthetic class. An alternative threshold mechanism was explored, with minimal impact on performance, and its results are reported in Appendix E.

Model	**Evaluation Alteration**					
	Auth.	DALLE3	SD1.X	SDXL	MJ 1/2	MJ 5/6
Single	-	97.70	98.30	97.57	99.07	99.25
Binary	94.85	99.64	**98.98**	99.22	99.60	99.74
6 Class	**97.39**	**99.76**	97.89	**99.30**	**99.91**	**99.97**

Table 3: Patch level recall for *Single*: five models trained on each synthetic dataset (*i.e.,* Table 2 diagonal), *Binary*: model trained with all synthetic datasets merged, *6 Class*: multi-class model trained for the recognition task. Best in bold.

Results in Table 3 show good performance from both the binary and the six-way classifiers on all synthetic datasets, better than single models, which means visual features of synthetic detectors are mutually beneficial for SID. In general, the six-way classifier outperforms all, with the only exception of one of the oldest and most distinct datasets (*i.e.,* SD1.X) (lowest generalization in Table 2).

4.3 Image Alteration Methods

Image transformations, while essential for storage optimization and transmission cost reduction, can significantly alter images and may be exploited by malicious actors to mask synthetic content. If SID models are not robust to these transformations, their utility in real-world scenarios becomes minimal. To evaluate this

robustness, we test the six-class model from the previous section under common transformations from the Albumentations library [1]: blur (*AdvancedBlur* and *GaussianBlur*), brightness and gamma alterations (*RandomBrightnessContrast* and *RandomGamma*) and *JPEG compression*, all using default parameters.

For a complete assessment, we train five multi-class models, each with a different transformation applied to its training set, and evaluate these alongside our original six-class model across all transformations and unaltered images. The results are presented in Table 4 using multi-class macro accuracy, where any misclassification between synthetic classes is counted as an error. This metric was selected instead of the previously used binary metrics, as binary classification consistently achieved over 99% accuracy, limiting its ability to distinguish model performance in a multi-class context.

Training Alteration	**Evaluation Alteration**						**Avg.**
	None	*Bright*	γ	*JPEG*	*ABlur*	*GBlur*	
None	**90.90**	86.66	90.60	90.19	81.56	54.73	82.44
Bright	91.28	**89.68**	91.13	90.10	84.61	63.55	85.06
γ	91.52	87.51	**91.30**	90.02	85.57	65.22	85.19
JPEG	87.82	83.15	87.79	**86.21**	78.42	55.29	79.78
ABlur	90.13	86.23	90.12	88.15	**88.04**	81.54	87.37
GBlur	88.94	84.02	88.65	87.37	86.78	**81.88**	86.27
Avg.	90.10	86.21	89.93	88.67	84.16	67.04	

Table 4: Patch-level accuracy (%) of six-class recognition models when trained on one alteration method and evaluated on all. In bold, performance on the alteration used for training. Last column: model average across all transformations. Bottom row: average performance of all models for each transformation.

Table 4 shows blur is the transformation that most impacts detector performance. *GaussianBlur*, which causes drops in accuracy of over 7 points, is also the hardest transformation in training, showing the lowest diagonal score. However, both blur-trained models achieve the highest cross-transformation accuracies, demonstrating effective generalization and making blur a valuable addition to the training process.

5 Deployment Experiments

To study the impact of deployment factors on generalization, we use `SuSy`, a multi-class model trained with the setup described in §4. Training data from §3.2 is augmented with all transformations from §4.3, each applied with a 20% chance. We explore generalization to new data sources (§5.1), patch-to-image prediction aggregation (§5.2) and resolution impacts (§5.3).

5.1 Generalization to Source

The *SuSy (Patch)* column of Table 5 shows evaluation results under disjoint sets of data (see §3.3). For authentic datasets, the model generalizes well to `Flickr30k` (91.81%), moderately to `GLDv2` (68.37%) and poorly to `In-the-wild` images (30.91%). Robust performance across these diverse datasets suggests minimal impact from potential content biases introduced in the training data.

Type	Data Source	SIG Model	Year	SuSy (Patch)	SuSy (Image)
Authentic	`Flickr30k`	None	2014	91.81	**94.48**
Authentic	`GLDv2`	None	2020	68.37	**71.80**
Authentic	`In-the-wild`	None	2024	**30.91**	27.27
Synthetic†	`Synthbuster`	SD 1.3	2022	88.56	**91.80**
Synthetic†	`Synthbuster`	SD 1.4	2022	88.50	**91.80**
Synthetic†	`Synthbuster`	MJ V5	2023	74.22	**78.40**
Synthetic†	`Synthbuster`	SD XL	2023	79.22	**83.80**
Synthetic†	`Synthbuster`	DALLE-3	2023	87.02	**92.50**
Synthetic‡	`Synthbuster`	Glide	2021	52.78	**53.20**
Synthetic‡	`Synthbuster`	SD 2	2022	68.32	**70.40**
Synthetic‡	`Synthbuster`	DALLE-2	2022	**24.50**	19.70
Synthetic‡	`Synthbuster`	Firefly	2023	53.04	**53.50**
Synthetic‡	`Authors`	SD 3	2024	91.51	**95.23**
Synthetic‡	`Authors`	FLUX.1	2024	94.37	**97.05**
Synthetic‡	`In-the-wild`	Unknown		90.51	**91.92**

Table 5: Recall at patch level and five-patch majority voting at image level for `SuSy`. Best in bold. †Generators seen during training. ‡Generators unseen during training.

The middle section of Table 5 shows datasets from generators seen during training but generated by different users. Although possible variations in SIG configurations, prompts and post-processing may introduce significant biases, `SuSy` achieves 74-88% recall across all cases.

For datasets from models unseen during training (bottom of Table 5), performance varies widely (24-94%). Model family inconsistently affects generalization: `SuSy` performs excellently on `SD3`, adequately on `SD2`, but poorly on `DALLE2`, despite training on versions of both generator families.

5.2 Image Decision Boundary

While SID models operate on image patches, real-world applications typically require whole-image predictions. To address this gap, we analyze the top five patches selected based on texture complexity (as described in §3.1).

We tested two aggregation strategies: majority voting of patch predictions and averaging patch logits before classification. Both improved over single-patch

performance, with majority voting consistently outperforming across datasets (results in the last column of Table 5). This approach further improved recall for high-performing datasets while providing minimal gains for poorly performing ones, revealing both advantages and limitations of decision boundary tuning.

5.3 Scale Generalization

Image resizing is a widespread image alteration that can alter or eliminate frequency artifacts that SID models rely on, potentially decreasing their performance. To assess the extent of this factor, we evaluate `SuSy` using images scaled at six different sizes (224 to 1440px). First, if the image is not already square, equal padding is added to the shorter dimension to center it. Then, the squared image is resized to the specified dimensions using bilinear interpolation, from which the evaluated patches are extracted. Using the evaluation datasets described in §3.3, which follow a diverse distribution of sizes (see Appendix G), we compute recall separately for authentic and synthetic classes at each scale. This approach allows us to monitor both accuracy and balance in detection across different resolutions. This experiment is reproduced in the benchmarking analysis of §6 for comparison with other SID models.

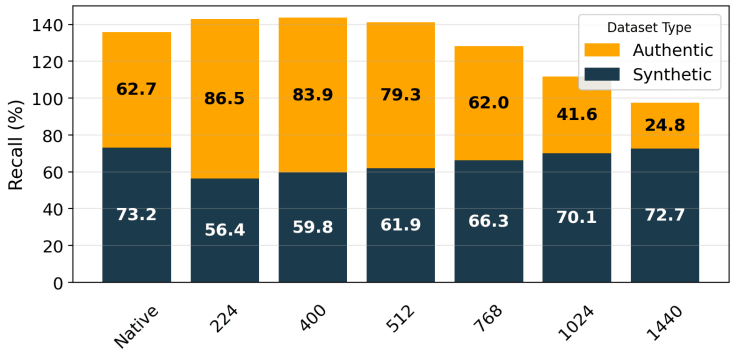

Fig. 3: Recall of `SuSy` under different scaling factors.

Figure 3 shows `SuSy` maintains stable performance at lower resolutions (224-512px). As resolution increases, predictions become increasingly biased toward the synthetic class. For consistent real-world performance, where images may have undergone prior resizing, we recommend including standardized rescaling in preprocessing pipelines.

5.4 Human Evaluators

To benchmark against human perception, we asked 10 social media users (aged 22–30) likely to be exposed to AI-generated content, to classify the `In-the-wild`

dataset. Images were presented in random order on identical IPS LCD displays under controlled lighting conditions. Participants were not informed about the class distribution. On average, volunteers took 15 minutes to classify all 210 images under no time constraints. Using our optimal setup from previous sections, `SuSy` outperformed the average human evaluator by 1.5%.

6 Benchmarking Experiments

To complete this study, we test the performance of ten different SID models (most available through SIDBench [34]). Table 6 showcases the six best-performing models (exceeding 140 combined recall points): LGrad [37], GramNet [29] and DIMD [23], which use CNNs as feature extractors, along with transformer-based models Rine [23], DeFake [11] and FatFormer [28]. Appendix F provides architectural details and results for the remaining tested detectors: CNNDetect [40], Dire [41], FreqDetect [18] and UnivFD [32].

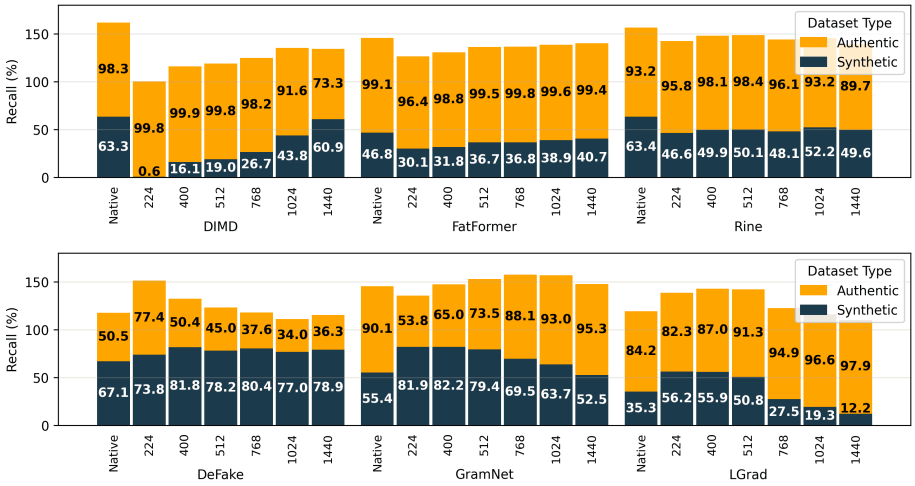

Fig. 4: Recall of state-of-the-art SID models under different scaling factors.

6.1 Rescaling

Given the crucial role of image scaling, detector performance is first assessed using the methodology from §5.3, providing insights on generalization under scale changes. Although DIMD, GramNet and LGrad were originally trained at 256×256, we standardize all evaluations at 224×224 patches across models to ensure consistency in processed information. To confirm the fairness of this approach, we performed a preliminary evaluation of these models at their native

256×256 resolution. The results, presented in Figure 5, show that the change in resolution does not fundamentally alter their performance characteristics.

The models in the top row of Figure 4 are highly sensitive to *any* scale modifications, with performance consistently deteriorating after rescaling (*i.e.,* optimal performance without resizing). This sensitivity creates a security vulnerability that malicious actors could exploit. Additionally, these models show bias toward the authentic class, with suboptimal synthetic image recall (63% for two models, while the third performs below random chance).

In contrast, detectors in the bottom row demonstrate resilience to *some* scale variations (*i.e.,* optimal performance includes resizing). DeFake and LGrad perform optimally at lower resolutions (224-512px), similar to `SuSy`, while GramNet excels at higher resolutions (768-1024px). Their optimal input resolution enhances resilience and enables deployment pipeline optimization. However, these models differ in prediction balance: DeFake excels in synthetic class detection, LGrad in authentic class identification, while GramNet and `SuSy` achieve more balanced performance across both categories.

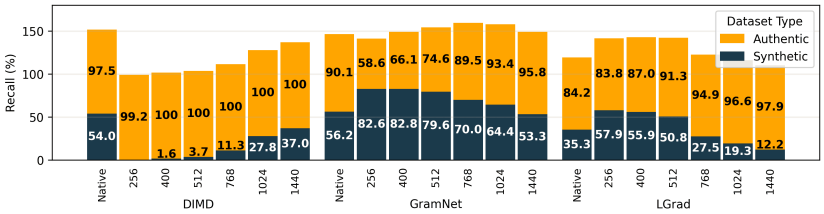

Fig. 5: Evaluation of DIMD, GramNet, and LGrad on 256×256 pixel patches, matching their original training resolution.

6.2 Optimal Model Generalization

The final experiment evaluates detector generalization across benchmarking datasets at optimal input scales. Results in Table 6 reveal a critical limitation: all detectors achieve less than 50% recall on at least four datasets, demonstrating that no universal detector exists. Performance metrics consistently favor the authentic class over synthetic, partly from optimal resolution selection and partly reflecting the more diverse and challenging synthetic distribution. A clear trade-off emerges: DeFake is simultaneously the best synthetic detector and worst authentic detector, while FatFormer shows the opposite pattern.

While DeFake leads in synthetic detection overall and DIMD excels in 8 of 17 synthetic datasets (including 6 of 7 StableDiffusion variants), both are highly sensitive to rescaling, making them unsuitable for deployment with uncontrolled inputs. DIMD's consistent performance across StableDiffusion models is exceptional, as detectors generally lack consistency across model families. Even when datasets use the same generative model, performance varies significantly

	SIG	Year	DIMD	FatFormer	Rine	DeFake	LGrad	SuSy	GramNet	Avg.
Resolution			Nat.	Nat.	Nat.	224	400	400	768	
Authentic Data										
Flickr30k	-	'14	99.92	**100.0**	99.54	93.62	99.82	94.76	99.94	**98.23**
COCO	-	'17	**100.0**	99.60	**100.0**	91.33	77.63	-	87.12	92.61
GLDv2	-	'20	96.54	99.92	77.42	78.32	98.66	82.62	**100.0**	90.50
In-the-wild	-	'24	96.69	**96.77**	95.87	46.28	71.90	74.38	65.29	78.17
Avg. Authentic			98.29	**99.07**	93.21	77.39	87.00	83.92	88.09	
Synthetic Data Sources										
Synthbuster	Glide	'21	6.10	68.10	83.60	**86.50**	53.50	53.30	68.80	59.99
mj-tti	MJ V1/V2	'22	2.87	55.29	14.57	**75.83**	42.60	-	63.58	42.46
Synthbuster	SD 1.3	'22	**100.0**	88.20	99.90	86.20	81.10	87.00	90.00	90.34
Synthbuster	SD 1.4	'22	**100.0**	88.00	99.60	87.20	81.30	87.10	90.70	**90.56**
diffusiondb	SD 1.X	'22	**99.92**	86.33	96.03	76.01	52.11	-	93.52	83.99
Synthbuster	SD 2	'22	**97.10**	47.20	85.80	39.80	53.80	42.30	75.50	63.07
Synthbuster	DALLE-2	'22	0.40	45.80	70.80	47.30	76.00	20.70	**93.10**	50.59
Synthbuster	MJ V5	'23	**98.10**	30.60	87.00	49.90	83.60	36.50	88.50	67.74
mj-images	MJ V5/V6	'23	**90.11**	5.16	31.28	75.85	8.27	-	15.56	37.71
Synthbuster	SDXL	'23	94.40	79.80	**97.60**	53.60	86.20	45.90	97.20	79.24
real.SDXL	SDXL	'23	**97.65**	28.64	82.17	91.82	69.77	-	75.61	74.28
Synthbuster	Firefly	'23	18.10	81.40	43.30	54.00	66.40	24.50	**83.00**	52.96
Synthbuster	DALLE-3	'23	0.00	0.00	2.00	**93.60**	35.00	84.30	30.40	35.04
dalle3-imgs	DALLE-3	'23	61.82	3.92	28.79	**81.21**	3.03	-	1.52	30.05
∗ SD3	SD 3	'24	**99.24**	59.02	85.05	89.42	70.74	78.44	89.03	81.56
∗ FLUX.1	Flux.1-dev	'24	62.89	23.08	54.72	81.43	63.92	85.40	**92.99**	66.35
In-the-wild	Unknown		47.47	5.00	16.16	**84.85**	22.22	71.72	32.32	39.96
Avg. Synthetic			63.30	46.80	63.43	**73.80**	55.86	59.76	69.49	

Table 6: Center-patch recall of detector models evaluated with their optimal input resize resolution (*Native* denotes no alteration). Best recall in bold and recalls below 50% underlined. Entries denoted by (-) for SuSy indicate training datasets excluded from evaluation.

— average recall differences between `DALLE3` versions is 27.63% across detectors, while `SDXL` variations average 24.35%. While detectors may generalize to source changes under specific conditions (see §5.1), this is not universal. Even in controlled scenarios like `Synthbuster`, where identical prompts are used, detection performance exhibits substantial variability.

Private models (`DALLE`, `Midjourney`, `Firefly`) present particular challenges, with detectors achieving only 45.22% average recall on closed SIG models compared to 76.60% on open ones. Even the *best*-performing closed dataset achieves only 7.75% better recall than the *worst* open one, stressing the importance of open science in advancing the field of SID.

The `In-the-wild` dataset, serving as a proxy for real-world conditions, reveals additional limitations. No tested detector exceeds 50% recall for *both* authentic and synthetic versions across all input resolutions. Only `SuSy` demonstrates robust performance with over 70% recall in both subsets, but specifically when operating at its optimal input resolution.

7 Conclusions

In a race equilibrium paradox, better generative models appear regularly, making the task harder for humans, while detectors trained on these newer generators are more reliable (see §4.1), keeping the race close.

The demand for detectors grows as society seeks to preserve social trust and digital rights while combating disinformation. Yet, these detectors must improve their generalization capabilities to be truly effective. In that regard, the main lesson from this work is: *never* assume generalization in SID. Results in Table 6 indicate even within datasets produced by the same generative model, detection performance may largely vary, as a result of software and hardware setups and user bias. Similarly, generalization should not be assumed on synthetic images produced by older, less realistic generators either, even if these synthetic samples seem more obvious to the human eye. As shown in Table 2, samples from these models are hard to generalize to (but *not* to train for) due to their stronger biases and distinct artifacts. In fact, even simple post-processing methods, like blur, can significantly reduce detector performance (see Table 4).

Image scale can dramatically affect the performance of most detectors, as well as the balance of their performance (*i.e., authentic* vs *synthetic*). Some detectors are highly sensitive to rescaling operations (see Figure 4), exposing a vulnerability to malicious inputs. At the same time, other SID models work optimally when applied to data that has been scaled to a certain size (see Figure 4). This can be used to tune data for its detection, boosting performance on deployment settings (see Table 6).

The final contribution of this work, beyond the released `SuSy`, code and datasets, is a list of policies for the SID field as a whole, including an ethical risk assessment. Our work emphasizes the importance of openness in generative AI. Results from Table 6 indicate open generative models can be more easily detected (+20% combined recall points on average). While we are far from a universal de-

tector (all detectors perform below random in some of our benchmarks), models trained for specific targets may be as good as humans at identifying synthetic content (see §5.4).

7.1 Ethical Risks

Image detection systems pose significant ethical concerns, primarily due to their inherent fallibility. These systems produce both false positives and negatives (see Table 6), potentially misidentifying authentic images as synthetic and vice versa. Such errors could infringe on digital rights and enable censorship. Therefore, human expert oversight is crucial when these systems are used in contexts affecting individual rights and their outputs should never serve as definitive evidence.

Additionally, model bias remains a critical challenge. Training datasets often contain inherent biases that can skew detection results (*e.g.,* rural landscapes could be tagged as synthetic more often than urban images). Thorough evaluation across all relevant demographic and contextual factors is essential before deployment. Furthermore, the datasets used for training may include samples with personal data. `COCO` contains images of real people and synthetic datasets used could include realistic depictions of specific individuals. However, given the training objective and parameter size of `SuSy`, it is highly unlikely that any such information could be encoded within the weights released in this work.

A final risk of releasing a SID model is dual use, as it can be used as a training objective for generative models (*e.g.,* adversarial training). To mitigate that, we add a specific clause in the terms of use of the model prohibiting such practice. Notice `SuSy` is not trained to be the best possible detector (not trained on all data) and should not be used *as is* in practice. We recommend any SID model produced for final use to be kept private, as long as its public release holds no special academic or social value.

7.2 Future Work

The results of this work point towards four research directions that could improve SID robustness and adaptability. The complementary strengths of different detector models indicate potential benefits from ensemble methods. Exploring training data scaling laws could reveal further insights into data requirements and generalization capabilities. Given the impact of input resolution, developing multi-resolution architectures could provide inherent resilience against scaling-based evasion attempts. Lastly, extending detection capabilities to video content is crucial to address the increasing quality of video generation models. These advancements are critical to ensure SID keeps pace in the ongoing race with SIG.

References

1. A. Buslaev, A. Parinov, E.K.V.I.I., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. ArXiv e-prints (2018)
2. Agency, E.E.: Greenhouse gas emission intensity of electricity generation in europe (2024), https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emission-intensity-of-1
3. Aghasanli, A., Kangin, D., Angelov, P.: Interpretable-through-prototypes deepfake detection for diffusion models. In: Proceedings of IEEE/CVF international conference on computer vision. pp. 467–474 (2023)
4. AI, S.: Stable diffusion 3 medium (2024), https://huggingface.co/stabilityai/stable-diffusion-3-medium
5. Bammey, Q.: Synthbuster: Towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing (2023)
6. Cazenavette, G., Sud, A., Leung, T., Usman, B.: Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10759–10769 (2024)
7. Chang, Y.M., Yeh, C., Chiu, W.C., Yu, N.: Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. arXiv preprint arXiv:2310.17419 (2023)
8. Chen, B., Zeng, J., Yang, J., Yang, R.: Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: Forty-first International Conference on Machine Learning (2024)
9. Coccomini, D.A., Esuli, A., Falchi, F., Gennaro, C., Amato, G.: Detecting images generated by diffusers. PeerJ Computer Science **10**, e2127 (2024)
10. Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., Verdoliva, L.: 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. pp. 973–982 (2023)
11. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
12. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L.: Raising the bar of ai-generated image detection with clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4356–4366 (2024)
13. Del Moral, P., Nowaczyk, S., Pashami, S.: Why is multiclass classification hard? IEEE Access **10**, 80448–80462 (2022)
14. Deng, Y., Deng, X., Duan, Y., Xu, M.: Diffusion-generated fake face detection by exploring wavelet domain forgery clues. In: 2023 International Conference on Wireless Communications and Signal Processing (WCSP). pp. 1–6. IEEE (2023)
15. DucHaiten: realisticsdxl (2023), https://huggingface.co/datasets/DucHaiten/DucHaiten-realistic-SDXL
16. ehristoforu: dalle-3-images (2024), https://huggingface.co/datasets/ehristoforu/dalle-3-images
17. ehristoforu: midjourney-images (2024), https://huggingface.co/datasets/ehristoforu/midjourney-images
18. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning. pp. 3247–3258. PMLR (2020)

19. Grommelt, P., Weiss, L., Pfreundt, F.J., Keuper, J.: Fake or jpeg? revealing common biases in generated image detection datasets. arXiv preprint arXiv:2403.17608 (2024)
20. Gustavosta: Stable-diffusion-prompts (2023), https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts
21. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics (6), 610–621 (1973)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Koutlis, C., Papadopoulos, S.: Leveraging representations from intermediate encoder-blocks for synthetic image detection. arXiv preprint arXiv:2402.19091 (2024)
24. Labs, B.F.: Flux.1-dev (2024), https://huggingface.co/black-forest-labs/FLUX.1-dev
25. Laurier, L., Giulietta, A., Octavia, A., Cleti, M.: The cat and mouse game: The ongoing arms race between diffusion models and detection methods. arXiv preprint arXiv:2410.18866 (2024)
26. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: Dit: Self-supervised pre-training for document image transformer. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3530–3539 (2022)
27. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
28. Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y.: Forgery-aware adaptive transformer for generalizable synthetic image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10770–10780 (2024)
29. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8060–8069 (2020)
30. Lorenz, P., Durall, R.L., Keuper, J.: Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 448–459 (2023)
31. López Cuena, E.: Super-resolution assessment and detection (06 2023), http://hdl.handle.net/2117/395959
32. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24480–24489 (2023)
33. Ricker, J., Damm, S., Holz, T., Fischer, A.: Towards the detection of diffusion model deepfakes. arXiv preprint arXiv:2210.14571 (2022)
34. Schinas, M., Papadopoulos, S.: Sidbench: A python framework for reliably assessing synthetic image detection methods. arXiv preprint arXiv:2404.18552 (2024)
35. Song, J., Ye, D., Zhang, Y.: Trinity detector: text-assisted and attention mechanisms based spectral fusion for diffusion generation image detection. arXiv preprint arXiv:2404.17254 (2024)
36. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space learning. arXiv preprint arXiv:2403.07240 (2024)

37. Tan, C., Zhao, Y., Wei, S., Gu, G., Wei, Y.: Learning on gradients: Generalized artifacts representation for gan-generated images detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12105–12114 (2023)
38. Turc, I., Nemade, G.: Midjourney user prompts & generated images (250k) (2022). https://doi.org/10.34740/KAGGLE/DS/2349267, https://www.kaggle.com/ds/2349267
39. Union, E.: Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2024), https://artificialintelligenceact.eu/
40. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
41. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22445–22455 (2023)
42. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. arXiv:2210.14896 [cs] (2022), https://arxiv.org/abs/2210.14896
43. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2575–2584 (2020)
44. Wu, H., Zhou, J., Zhang, S.: Generalizable synthetic image detection via language-guided contrastive learning. arXiv preprint arXiv:2305.13800 (2023)
45. Xi, Z., Huang, W., Wei, K., Luo, W., Zheng, P.: Ai-generated image detection using a cross-attention enhanced dual-stream network. In: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1463–1470. IEEE (2023)
46. Xu, Q., Wang, H., Meng, L., Mi, Z., Yuan, J., Yan, H.: Exposing fake images generated by text-to-image diffusion models. Pattern Recognition Letters **176**, 76–82 (2023)
47. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
48. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE international workshop on information forensics and security (WIFS). pp. 1–6. IEEE (2019)
49. Zhong, N., Xu, Y., Li, S., Qian, Z., Zhang, X.: Patchcraft: Exploring texture patch for efficient ai-generated image detection. arXiv preprint arXiv:2311.12397 pp. 1–18 (2024)
50. Zhu, M., Chen, H., Huang, M., Li, W., Hu, H., Hu, J., Wang, Y.: Gendet: Towards good generalizations for ai-generated image detection. arXiv preprint arXiv:2312.08880 (2023)
51. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. Advances in Neural Information Processing Systems **36** (2024)