Synthetic Images at MediaEval 2025: Advancing detection of generative AI in real-world online images

Olga Papadopoulou[1,*,†], Manos Schinas[1,†], Riccardo Corvi[2,†], Dimitrios Karageorgiou[1,3], Christos Koutlis[1], Fabrizio Guillaro[2], Efstratios Gavves[3], Hannes Mareen[4], Luisa Verdoliva[2] and Symeon Papadopoulos[1]

[1] *Information Technologies Institute (ITI) @ CERTH, Greece*
[2] *University Federico II of Naples, Italy*
[3] *University of Amsterdam, The Netherlands*
[4] *IDLab, Ghent University — imec, Belgium*

Abstract

The task *Synthetic Images: Advancing detection of generative AI used in real-world online images* at MediaEval2025 focuses on the challenge of developing AI models capable of detecting synthetic images and identifying the specific regions in the images that have been manipulated or synthesized. Participants face two subtasks: 1. *Synthetic image detection* where AI methods classify images as real or synthetic and 2. *Manipulated region localization* where the AI method needs to localize the specific region(s) in an image that have been generated or modified. In addition to these tasks, method robustness is evaluated by assessing their ability to retain detection performance under real-world transformations, such as compression, resizing, and cropping, commonly found on social media. Results show that while several teams achieved promising outcomes in the constrained detection task, generalizable synthetic image detection and manipulation localization in fully regenerated images remain challenging. A dedicated GitHub repository accompanies the task, providing all necessary datasets, evaluation process and a public leaderboard for benchmarking participant submissions[1].

1. Introduction

Synthetic media is increasingly used in creative industries, social media, and, alarmingly, in misinformation campaigns. Synthetic image detection is vital for combating the misuse of AI-generated content, ensuring trust in visual media, and upholding ethical standards in digital communication. Recent advancements in synthetic image classification have shown promise [1, 2], yet challenges persist [3] due to evolving generative models and post-processing techniques such as image recompression that obfuscate traces of generative AI.

In addition, the advent of text-guided generative AI tools enables realistic inpainting [4, 5], including adding new objects or modifying specific regions. Traditionally, manipulations were performed without any sophisticated AI method and required special skills and effort with tools like Photoshop. Images manipulated through a manual process are commonly referred to as cheapfakes. The term deepfake, instead, more commonly refers to modifications performed

[1] https://github.com/mever-team/mediaeval2025-sid

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

[*] Corresponding author.

[†] These authors contributed equally.

olgapapa@iti.gr (O. Papadopoulou); manosetro@gmail.com (M. Schinas); riccardo.corvi@unina.it (R. Corvi); dkarageo@iti.gr (D. Karageorgiou); ckoutlis@iti.gr (C. Koutlis); fabrizio.guillaro@unina.it (F. Guillaro); e.gavves@uva.nl (E. Gavves); hannes.mareen@ugent.be (H. Mareen); verdoliv@unina.it (L. Verdoliva); papadop@iti.gr (S. Papadopoulos)

CEUR Workshop Proceedings (CEUR-WS.org)

by AI tools, which allow for image manipulations in ways that retain semantic consistency. However, beyond the distinction between cheapfakes and deepfakes, a further distinction can be made between spliced and fully regenerated deepfakes, stemming from the different ways in which inpainting methods perform the content manipulation [5].

This challenge aims to bring together researchers and practitioners to push the boundaries of synthetic image detection. By focusing on both fully synthetic and partially modified images, this challenge highlights the complex nature of synthetic content in real-world scenarios and seeks solutions that are robust, generalizable, and explainable.

2. Task Description

The task is organized into two subtasks. In the following sections, the task description, provided data, and evaluation process are presented for each subtask.

2.1. Subtask A: Real vs. Synthetic Image Detection

The challenge is to develop approaches that classify whether a given image is real or synthetic. The subtask is structured into two runs: i) Constrained Run: Use only the official training dataset, ii) Open Run: Use any data, including external or self-generated synthetic images. The goal is to build and evaluate models that can accurately detect whether a given image is synthetic or real, especially under the challenging conditions posed by images in the wild.

Dataset The dataset comprises three parts: a training set (5,000 real and 5,000 synthetic in-the-wild images [6, 1]), a validation set for development (10,000 labelled real/synthetic (balanced) images), and a test set of 10,000 images without labels used by the organizers for the final evaluation. The training and validation data combine well-known synthetic image detection datasets with newly generated synthetic content and real-world samples from social media to ensure “in-the-wild” evaluation. Specifically, real images are collected from open datasets such as LAION [7] and RAISE [8], ensuring authenticity and diversity while synthetic images are produced by a variety of generative models, from GAN-based (StyleGAN2 [9], ProGAN [10], GigaGAN [11]) to diffusion models (Stable Diffusion [12], MidJourney, DALL·E 3 [13], Adobe Firefly). Finally, to ensure real-world variations, synthetic images are further transformed through compression, cropping, and other edits to simulate realistic online conditions. All data are curated under open or permissive licenses.

Evaluation metrics For evaluating synthetic image detection, we followed the SIDBench framework [14] which provides a comprehensive set of metrics to assess model robustness in real-world scenarios. While multiple indicators are reported (Accuracy, Precision, Recall, F1-Score, AUC, AP, and EER), F1-Score is used as the main ranking criterion, as it best reflects the balance between false alarms and missed detections which is crucial for reliable deployment in the wild.

2.2. Subtask B: Manipulated Region Localization

Subtask B focuses on the localization of AI-generated image manipulations, on both spliced and fully regenerated deepfakes. With the rapid advancement of Generative AI, modern tools such as diffusion-based inpainting systems now enable users to perform complex and realistic image edits through simple text prompts. These tools can introduce subtle or extensive modifications from inserting or replacing objects to regenerating entire scenes while maintaining semantic coherence with the original image.

Dataset The dataset includes both real and AI-manipulated images, designed to capture a broad range of manipulation types and complexities. It comprises spliced images, where newly generated regions are embedded in original content, and fully regenerated images, in which the entire scene is recreated even if only part of it has been semantically altered. This composition allows participants to evaluate their systems under both local and global manipulation conditions, reflecting the challenges encountered in real-world detection scenarios.

Participants are required to use the TGIF dataset [4] for training, and may use the dataset for validation. The TGIF dataset is split into two types of manipulated images: spliced (sp), and fully regenerated (fr). The spliced images were made using Adobe Photoshop (ps) and Stable Diffusion 2 (sd2), whereas the fully regenerated images were made using Stable Diffusion 2 and Stable Diffusion XL (sdxl).

The validation and test data originate from the SAGI-D dataset [5], created using the homonymous framework for generating semantically aligned text-guided image inpaintings. This data comprise alterations generated using AI methods that rely on splicing (Inpaint-Anything, Remove-Anything), full regeneration (ControlNet) or both (BrushNet, PowerPaint, HD-Painter). The validation dataset consists of 3,227 pristine images sourced from COCO (1,950) and RAISE (1,227) datasets as well as 6,212 manipulated images. The test dataset consists of real images from the COCO (2,922), RAISE (580) and OpenImages (5,570) datasets and a total of 10,178 manipulated images.

Evaluation metric For image-level detection, performance is assessed using the F1 score and the Area Under the ROC Curve (AUC), which respectively measure threshold-dependent classification performance and threshold-agnostic discriminative ability of a detector. For localization, the Intersection over Union (IoU) metric is employed to evaluate the overlap between predicted and ground-truth manipulated regions. We compute the IoU both with the predicted mask and with its inverted version, and then pick the higher of the two. This adjustment accounts for some localization methods that correctly separate the two areas, but misclassify which one is manipulated and which is pristine. Together, these metrics enable comprehensive evaluation of approaches' ability to both detect and precisely localize AI-generated manipulations.

3. Results and Analysis

As baselines for synthetic image detection, we employed three well-established methods: UniFD [15], RINE[2] and BFree [16]. UniFD employs a CLIP-trained Vision Transformer (ViT) as its backbone to detect synthetic images while RINE builds upon this by leveraging intermediate ViT layers to capture low-level visual artifacts. BFree adopts a bias-free training paradigm with diffusion images. We used UniFD trained on GAN data and RINE trained on diffusion data from the task while for BFree we kept the original model. As shown in Table 1, all methods perform poorly across metrics, highlighting limited generalization to real-world online images when models are trained on narrow, domain-specific datasets. BFree achieves higher performance, reaching an F1 score of 0.740 but its performance is still far from perfect, as indicated by the relatively lower recall value (0.635). A second RINE variant is trained on the TWIGMA dataset [17], that to the best of our knowledge, is the only large-scale dataset composed of in-the-wild synthetic images collected from the web. This exhibits improved performance, suggesting that more diverse and realistic training data can enhance generalization.

Model	Accuracy	Precision	Recall	F1
UniFD [15]	0.468 / 0.500	0.446 / 0.500	0.265 / 1.000	0.333 / 0.667
RINE [2]	0.537 / 0.543	0.521 / 0.524	**0.935** / 0.929	0.669 / 0.670
BFree [16]	0.777 / 0.773	**0.887** / 0.771	0.635 / 0.777	0.740 / 0.774
RINE(TWIGMA)	**0.789** / 0.801	0.751 / 0.787	0.865 / 0.827	**0.803**/ 0.806

Table 1
Performance of SID methods. We report Acc., Prec., Rec., F1 scores using both the validation-calibrated threshold and the optimal threshold, the latter serving as an upper performance bound.

As baselines to evaluate the complexity of distinguishing between both versions of deepfakes we use two state-of-the-art methods: TruFor [18] and DeCLIP [19]. TruFor is an image forensics approach developed for the detection and localization of local manipulations. It relies on two main components, one for the extraction of a low-level noise residual map (called Noiseprint++) as well as high-level RGB features, and another for the estimation of a confidence map that is used to compute a more informed and reliable image-level detection score. DeCLIP, instead, tries to explicitly tackle the case where the manipulations are semantically localized but forensically global, leveraging CLIP features and training a decoder to identify areas where an image might be manipulated.

Table 2 presents image-level detection performance. DeCLIP performs better than TruFor in terms of F1 while worse in terms of AUC. The latter, in turn, performs much better on spliced images while performing worse on the fully regenerated images. With respect to localization performance in Table 2, DeCLIP, despite being much newer and designed to challenge these mixed scenario, struggles to detect the localized areas, while TruFor is capable of perfectly localizing the spliced images. Although the approach was trained on cheapfakes, the disruption in the noise pattern is also introduced by AI-based inpainting, which leads to correct localizations. However, on fully regenerated images the performance drops quite significantly. This happens because artifacts of fully regenerated ones are consistent across the image.

		Detection (F1/AUC %)		Localization (IoU)	
Method	**Type**	**TruFor**	**DeCLIP**	**TruFor**	**DeCLIP**
Inpaint-Anything	SP	0.412 / 92.59	0.645 / 57.20	0.842	0.401
ControlNet	FR	0.219 / 52.79	0.654 / 59.14	0.468	0.430
BrushNet	SP/FR	0.412 / 58.81	0.657 / 61.14	0.678	0.530
PowerPaint	SP/FR	0.313 / 58.33	0.665 / 61.65	0.625	0.530
HD-Painter	SP/FR	0.200 / 55.87	0.661 / 61.39	0.439	0.486
Remove-Anything	SP	0.638 / 76.51	0.651 / 51.69	0.749	0.374
AVG		0.469 / 64.95	0.676 / 57.45	0.627	0.470

Table 2
Performance of state-of-the-art methods on the SAGI [5] dataset for both detection (F1/AUC) and localization (IoU). The sets include spliced (SP) and fully regenerated (FR) images.

4. Acknowledgments

The task organization is supported by the Horizon Europe AI-CODE and vera.ai projects that focus on the development of AI tools for supporting media professionals in their verification and fact-checking activities.

Declaration on Generative AI

ChatGPT 5 was used for Grammar and spelling check of this paper. The authors reviewed and edited the resulting content as needed and take full responsibility for the publication's content.

References

[1] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic images generated by diffusion models, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[2] C. Koutlis, S. Papadopoulos, Leveraging representations from intermediate encoder-blocks for synthetic image detection, in: European Conference on Computer Vision (ECCV), Springer, 2024, pp. 394–411.

[3] D. Karageogiou, Q. Bammey, V. Porcellini, B. Goupil, D. Teyssou, S. Papadopoulos, Evolution of detection performance throughout the online lifespan of synthetic images, in: European Conference on Computer Vision (ECCV) Workshops, Springer, 2024, pp. 400–417.

[4] H. Mareen, D. Karageorgiou, G. Van Wallendael, P. Lambert, S. Papadopoulos, TGIF: Text-guided inpainting forgery dataset, in: Int. Workshop on Information Forensics and Security (WIFS), 2024.

[5] P. Giakoumoglou, D. Karageorgiou, S. Papadopoulos, P. C. Petrantonakis, SAGI: Semantically aligned and uncertainty guided ai image inpainting, arXiv preprint arXiv:2502.06593 (2025).

[6] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. 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 (CVPR), 2020.

[7] C. Schuhmann, R. Kaczmarczyk, A. Komatsuzaki, A. Katta, R. Vencu, R. Beaumont, J. Jitsev, T. Coombes, C. Mullis, Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, in: NeurIPS Workshop Datacentric AI, FZJ-2022-00923, Jülich Supercomputing Center, 2021.

[8] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, G. Boato, Raise: a raw images dataset for digital image forensics, in: Proceedings of the 6th ACM Multimedia Systems Conference, 2015, pp. 219–224.

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and Improving the Image Quality of StyleGAN, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119.

[10] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations (ICLR), 2018.

[11] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, T. Park, Scaling up gans for text-to-image synthesis, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10124–10134.

[12] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, in: International Conference on Learning Representations (ICLR), 2024.

[13] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al., Improving image generation with better captions, Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf 2 (2023) 8.

[14] M. Schinas, S. Papadopoulos, SIDBench: A python framework for reliably assessing synthetic image detection methods, in: Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation, 2024, pp. 55–64.

[15] U. Ojha, Y. Li, Y. J. Lee, Towards universal fake image detectors that generalize across generative models, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24480–24489.

[16] F. Guillaro, G. Zingarini, B. Usman, A. Sud, D. Cozzolino, L. Verdoliva, A Bias-Free Training Paradigm for More General AI-generated Image Detection, in: IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18685–18694.

[17] Y. Chen, J. Y. Zou, TWIGMA: A dataset of AI-Generated Images with Metadata From Twitter, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems (NeurIPS), volume 36, 2023, pp. 37748–37760.

[18] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20606–20615.

[19] S. Smeu, E. Oneata, D. Oneata, DeCLIP: Decoding clip representations for deepfake localization, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 149–159.