Deep Learning-Based Detection and Localization of Generative AI Manipulations in Real-World Images

Konda Nitish[†], Jayasree R[†], Mirunalini P* and Ponsubash Raj R

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Abstract

With the rapid advancement of generative AI models, distinguishing authentic images from synthetically generated or manipulated ones has become a critical challenge for ensuring media authenticity. This research work presents a deep learning-based approach for the MediaEval 2025 Synthetic Image Detection challenge, which involves two subtasks: classification of real and synthetic images and localization of manipulated regions in spliced images. For Task A, the proposed system employs a ConvNeXt-based convolutional neural network to classify images as real or synthetic. For Task B, the framework utilizes a U-Net architecture with a ResNet-18 encoder to predict pixel-level manipulation masks. These masks are further used to generate binary localization maps and compute global manipulation probabilities. Task A achieves an F1 score of 0.544 (constrained run), while Task B achieves an F1 score of 0.35 and an IoU of 0.30 for localization, along with an F1 of 0.36 for detection across all sources. Experimental results indicate that the proposed pipeline provides reasonable performance in both classification and localization, offering a solid baseline for further improvement in real-world image manipulation detection and analysis.

1. Introduction

With the advent of sophisticated generative models in today's digital age, it has become even more challenging to distinguish between authentic and artificially generated images. Those synthesized or manipulated images are readily available online, causing concern regarding misinformation and veracity. Identifying these images and recognizing which areas within them were modified are essential to maintaining trust in digital media. The Synthetic Image Detection (SID) task [1] aims to solve this issue by trying to achieve two goals : detection or classification of whether the image is synthetic or real, and localization of the exact areas that have been generated or altered.

2. Related Work

The detection of synthetic images generated by generative AI models has drawn significant research attention due to the rising prevalence of realistic fake media. Recent works have focused on improving detection robustness and generalization. Koutlis [2] leverage intermediate encoder-block representations, while Guillaro [3] propose a bias-free training paradigm for more general AI-generated image detection. Karageorgiou [4] introduce any-resolution detection via spectral learning. Li [5] developed the SAFE detector, leveraging diverse augmentations for better generalization, and Ojha [6] proposed a CLIP-based universal detector that generalizes across

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

*Corresponding author.

[†]These authors contributed equally.

nitish2370015@ssn.edu.in (K. Nitish); jayasree2370002@ssn.edu.in (J. R); miruna@ssn.edu.in (M. P); ponsubashraj2370043@ssn.edu.in (P. R. R)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

multiple generative models. Earlier, Wang [7] highlighted the importance of data augmentation for improving detection robustness.

In addition to classifying synthetic images, localizing the specific manipulated regions is critical. Bappy [8] introduce a multi-task deep learning framework that combines patch-level classification and pixel-wise segmentation, exploiting the spatial structure of manipulations to accurately localize altered regions. More recently, Guillaro [9] proposed Trufor, which leverages multiple complementary cues for trustworthy image forgery detection and localization.

The proposed framework integrates robust synthetic image classification with precise manipulated region localization, enabling accurate detection and pinpointing of manipulations in realistic scenarios for practical forensic use.

3. Approach

3.1. Data Description

The first task focuses on real versus synthetic image classification using combined datasets from Wang [10] and Corvi [11], which include real images from COCO and LSUN alongside synthetic ones generated by StyleGAN2, BigGAN, ProGAN, and Latent Diffusion Models (LDMs). The training dataset is balanced between real and fake samples, the validation set contains 10,000 labeled images (5,000 real and 5,000 synthetic), and the test set includes 10,000 unlabeled images from similar sources. For open run, in addition to the official MediaEval dataset, we augmented training by incorporating the publicly available CIFAKE dataset [12], which contains a diverse collection of real and AI-generated synthetic images.

The second task addresses manipulated region localization using the TGIF dataset [13], which includes authentic and manipulated images across six categories, with manipulations created through splicing or full regeneration using tools like Photoshop, Stable Diffusion 2, and Stable Diffusion XL. Each image is paired with a binary ground-truth mask marking edited areas, while validation data from COCO and RAISE include both original and manipulated samples generated by seven inpainting or diffusion-based methods (e.g., brushnet, controlnet, hdpainter, powerpaint) with corresponding pixel-level masks.

3.2. Methodology

The proposed framework consists of two parallel pipelines corresponding to Task A (real vs. synthetic classification) and Task B (manipulated region localization). Figures 1 and 2 illustrate the architectures for Task A and Task B, respectively. The input images are first preprocessed by resizing and normalization to standardize the data before being passed to the models.

For the first task, a convolutional neural network (CNN)-based classifier built using the ConvNeXt architecture is employed. The model is trained to distinguish between real and manipulated images by learning discriminative visual patterns. The ConvNeXt model is fine-tuned on the official training set using balanced batches of authentic and tampered images. The final classification layer outputs a single sigmoid-activated probability representing the likelihood of manipulation.

For the second task, a U-Net-based encoder–decoder architecture is used to achieve pixel-wise localization of manipulated regions in images. The encoder, built on a ResNet-18 backbone, extracts multi-scale contextual features, while skip connections preserve spatial information by directly linking encoder outputs to corresponding decoder stages. The decoder upsamples these features, reconstructing fine spatial details necessary for precise segmentation. At the final stage, a 1×1 convolutional layer is applied to the decoder's output, projecting the multi-channel

feature map down to two channels—each representing the raw score (logit) for the "real" and "manipulated" classes at every pixel. These logits are converted to per-pixel probabilities via a softmax activation, producing a manipulation probability mask. The global manipulation probability is obtained by averaging all pixel-level probabilities of the manipulated class across the predicted mask. This mean score represents the overall likelihood of manipulation which is compared with the threshold for final classification. During training, each input image is paired with its binary ground-truth mask, and the model is optimized using a combination of Binary Cross-Entropy and Dice loss to address class imbalance and enhance segmentation accuracy.

Both models are trained using the Adam optimizer with learning rate scheduling and early stopping based on validation performance. Cross-dataset validation is performed to ensure generalization, training on one manipulation type and testing on another.

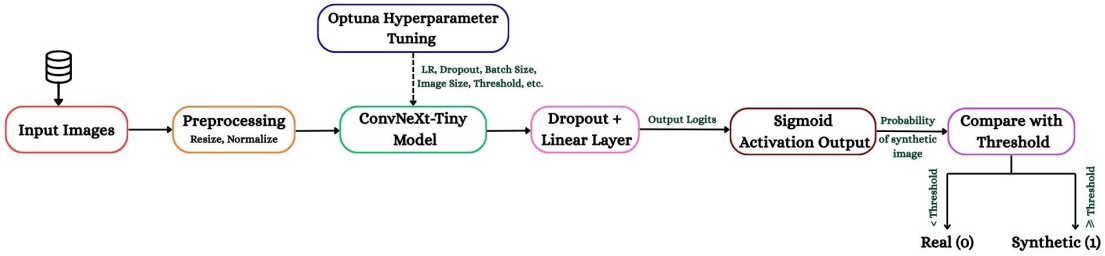

Figure 1: Task A: Real vs. Synthetic Classification

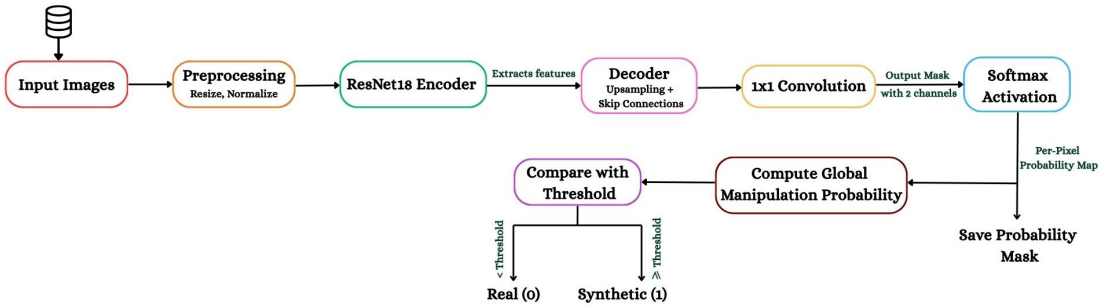

Figure 2: Task B: Manipulated Region Localization

4. Results and Analysis

For Task A, the evaluation metric is the F1 score. For Task B, evaluation is performed in two stages: detection of manipulated images using balanced accuracy and F1 score, and localization of manipulated regions using F1 score and Intersection over Union (IoU). The results correspond to the average performance across all image sources. Table 1 summarizes the results on the official test set (10,000 images) for Task A, showing constrained and open runs.

The results indicate that the constrained run achieves a higher F1 score compared to the open run, while the open run achieves slightly better accuracy. The inclusion of additional diverse samples in the open run might have introduced a distribution mismatch that reduced performance. This suggests that the model performs better when evaluated on data similar to its training distribution.

Table 1
Task A Results on Official Test Data

Run Type	Accuracy	F1 Score
Constrained	0.4589	0.5436
Open	0.5047	0.3083

Table 2 presents the image-level detection results for Task B, showing balanced accuracy and F1 score averaged across all methods for each source type, as well as the overall average.

Table 2
Task B Detection Results (Image-level)

Source	Balanced Accuracy	F1 Score
OpenImages	0.5155	0.3476
COCO	0.5033	0.3460
Raise	0.4427	0.4416
Average	0.5102	0.3624

Table 3 reports the pixel-level localization results for Task B, showing F1 score and IoU averaged across all methods for each source type, as well as the overall average.

Table 3
Task B Localization Results (Pixel-level)

Source	F1 Score	IoU
Raise	0.4880	0.4248
COCO	0.3057	0.2517
OpenImages	0.3456	0.3016
Average	0.3575	0.3075

Overall, the results indicate that Task A achieves moderate F1 scores under both constrained and open settings, with constrained runs being more precise. For Task B, image-level detection achieves balanced accuracy around 0.51 and F1 score around 0.36, while pixel-level localization achieves an average F1 score of 0.36 and IoU of 0.31. These results highlight that both detection and precise localization of manipulated content remain challenging, and there is substantial scope for improving performance in realistic manipulation detection and segmentation.

5. Discussion and Outlook

Detecting and localizing manipulated content in images is a challenging task. Despite advances in deep learning architectures, achieving robust performance across diverse manipulation types and image sources is still difficult. There is substantial scope for improvement in both detection and precise localization, particularly for realistic and complex manipulations.

Future work may benefit from exploring more sophisticated model architectures, multi-scale feature extraction [14], and attention mechanisms, as well as leveraging larger and more diverse datasets. Techniques such as self-supervised pretraining, domain adaptation [15], or integration of multi-modal information could further enhance the generalization and accuracy of manipulation detection and segmentation systems. Continued research in this area is essential to develop more reliable tools for ensuring the authenticity of visual content.

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4 and Gramby for grammar and spelling checks. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] O. Papadopoulou, M. Schinas, R. Corvi, D. Karageorgiou, C. Koutlis, F. Guillaro, E. Gavves, H. Mareen, L. Verdoliva, S. Papadopoulos, Synthetic images at mediaeval 2025: Advancing detection of generative ai in real-world online images, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025. 25–26 October 2025.

[2] C. Koutlis, S. Papadopoulos, Leveraging representations from intermediate encoder-blocks for synthetic image detection, 2024. URL: https://arxiv.org/abs/2402.19091. arXiv:2402.19091.

[3] F. Guillaro, G. Zingarini, B. Usman, A. Sud, D. Cozzolino, L. Verdoliva, A bias-free training paradigm for more general ai-generated image detection, 2025. URL: https://arxiv.org/abs/2412.17671. arXiv:2412.17671.

[4] D. Karageorgiou, S. Papadopoulos, I. Kompatsiaris, E. Gavves, Any-resolution ai-generated image detection by spectral learning, 2025. URL: https://arxiv.org/abs/2411.19417. arXiv:2411.19417.

[5] A. Li, et al., Improving synthetic image detection towards generalization, arXiv preprint arXiv:2408.06741 (2024). URL: https://arxiv.org/abs/2408.06741.

[6] U. Ojha, Y. Li, Y. J. Lee, Towards universal fake image detectors that generalize across generative models, 2024. URL: https://arxiv.org/abs/2302.10174. arXiv:2302.10174.

[7] S. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot... for now, CoRR abs/1912.11035 (2019). URL: http://arxiv.org/abs/1912.11035. arXiv:1912.11035.

[8] A. Bappy, et al., Exploiting spatial structure for localizing manipulated image regions, in: ICCV 2019, 2019. URL: https://vcg.ece.ucr.edu/sites/default/files/2019-02/iccv_jawad.pdf.

[9] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization, 2023. URL: https://arxiv.org/abs/2212.10957. arXiv:2212.10957.

[10] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot...for now, in: CVPR, 2020.

[11] 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. doi:10.1109/ICASSP49357.2023.10095167.

[12] J. J. Bird, A. Lotfi, Cifake: Image classification and explainable identification of ai-generated synthetic images, IEEE Access 12 (2024) 15642–15650. doi:10.1109/ACCESS.2024.3356122.

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

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[15] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7167–7176.