Examining Difficulties with CNNs and Segmentation Models for Real-vs. Synthetic Image Detection

Amrit Gopinath[1], Amudhan A[1], Adithya Sivakumar[1] and Vijayalakshmi[1]

[1]*Sri Sivasubramaniya Nadar College of Engineering, India*

Abstract

We participated in the MediaEval 2025 Synthetic Image Detection task, which involved real-vs-synthetic image classification (Task A) and manipulation localization (Task B). Our CNN-based classification model did not achieve high accuracy (Task A F1: 0.0627) because the results attained revealed a strong bias toward real images, highlighting the difficulty of generalizing across diverse synthetic sources. For Task B, our U-Net model achieved average F1 scores of 0.357 on segmentation and 0.362 on detection. The study identifies determinants of the performance shortfall, discusses dataset bias, and presents strategic insights for establishing a more robust model in future work.

1. Introduction

Given the widespread use of synthetic images and image manipulations, considerable vulnerabilities are introduced across various applications. This necessitates that the detection of these forgeries is a vital capability, especially in fields like social media platform moderation and digital forensic analysis.

Task A involves the fundamental challenge of classifying real images against synthetic images produced by prominent GANs (such as ProGAN, StyleGAN2, and BigGAN) under practical circumstances. **Task B** is designed to precisely identify the altered sections within both original and manipulated media, including images spliced using tools like Photoshop or SD2, as well as those fully generated by the SD2 or SDXL models.

We developed CNN-based pipelines for classification and U-Net architectures for segmentation. Our contributions include:

- Analyzing CNN performance for real-vs-synthetic image detection, which revealed a strong class bias and subsequent generalization difficulties.
- Achieving pixel-level localization of manipulated areas, a capability that shows strong generalization across a wide variety of forgery types.
- Conducting a comprehensive evaluation on the official validation sets to firmly establish the effectiveness of our proposed method.

Even after significant research investment, models often develop dataset-specific biases, preventing them from reliably generalizing across different or previously unobserved generative sources. We aimed to explore such challenges using convolutional architectures for classification and segmentation tasks. The outcomes reveal limitations in the path chosen and directions that can be followed in order to improve generalization in this regard.

MediaEval'25, 25–26 October 2025, Dublin, Ireland and Online

amrit2410182@ssn.edu.in (A. Gopinath); amudhan2410622@ssn.edu.in (A. A); adithya2410402@ssn.edu.in (A. Sivakumar); vijayalakshmia@ssn.edu.in (Vijayalakshmi)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Work

This paper addresses the Synthetic Images task as described in the task overview paper [1]. The detection of synthetic images has been studied with reference to GAN and diffusion model generation. CNN-based classifiers, such as ResNet and DenseNet, have shown strong performance on real-vs-synthetic tasks [2]. In order to improve the generalization of in-the-wild datasets, data augmentation and transfer learning are key [3].

Pixel-level manipulation localization typically uses encoder-decoder architectures, like U-Net [4, 5]. The TGIF dataset provides diverse manipulation types, making it challenging for these approaches [3].

3. Approach

3.1. Task A — Real vs. Synthetic Image Detection

For Task A, we use pretrained DenseNet121 fine-tuned on 5,000 real images and 5,000 synthetic images (10,000 total). The data augmentation includes horizontal flips, rotations, and color jitter.

A post-evaluation review showed that the model's training was heavily biased toward real-image features. The consequence was that the classifier defaulted to labeling almost every image as real, which explains the poor performance metric (an F1-score of 0.0627). This behavior points to a clear deficiency in either the synthetic feature representation or the basis of the current augmentation approach.

3.2. Task B — Manipulation Localization

The approach for Task B started by organizing the images into real and fake categories (covering ps-sp, sd2-sp, sd2-fr, sdxl-fr), all of which included a corresponding binary mask. Images then underwent necessary preprocessing, such as resizing, ImageNet normalization, and routine augmentations.

- **Classification:** CNN models were utilized to make the overall prediction of whether an image was authentic or manipulated.
- **Segmentation:** The U-Net architecture was used to determine the precise pixel location of the manipulation via masks.
- **Training:** involved the typical optimization cycle (forward pass, loss, backpropagation, and updating weights). We assessed the models using pixel accuracy, mIoU, and the F1-score.

For this submission, we primarily used U-Net for segmentation tasks. Global classification decisions were made by capping the proportion of pixels predicted as manipulated in the segmentation mask. Segmentation outputs are used for pixel-level evaluation, and the global classification for Task B is inferred from the mask coverage.

4. Results and Analysis

4.1. Task A – Real vs. Synthetic Classification

The model achieved an F1-score of 0.0627 and an accuracy of 0.5005, which shows poor generalization and strong bias toward the real class of images.

Table 1
Performance Metrics for Task A

Metric	Value
Accuracy	0.5005
Precision	0.5076
Recall	0.0334
F1	0.0627
ROC AUC	0.5512
Average Precision	0.5278

4.2. Task B – Manipulation Detection & Segmentation

Table 2 shows the results of our detection approach, reporting balanced accuracy, AUC, and F1_th for the top-performing methods and overall averages.

Table 2
Detection Metrics for Task B (Top 3 methods overall, F1_th values)

Source	Method	Balanced Accuracy	AUC	F1_th
RAISE	hdpainter	0.4724	0.4719	0.4658
RAISE	removeanything	0.4844	0.4893	0.4500
OpenImages	removeanything	0.5480	0.5477	0.4069
AVG	AVG	0.5102	0.5123	0.3624

Similarly, Table 3 shows the segmentation performance of our approach, evaluated through IoU and F1 metrics using the achieved threshold (F1_th) values.

Table 3
Segmentation Metrics for Task B (F1_th values)

Folder	Method	F1_th	IoU
RAISE	mixed	0.5745	0.5075
COCO	mixed	0.4615	0.3733
OpenImages	mixed	0.3856	0.3284
AVG	–	0.3570	0.3075

Task B segmentation resulted in an average F1_th score of 0.357 and IoU of 0.3075, with the highest F1_th score of 0.649 on RAISE-Mixed.

5. Conclusions

In this paper, we analyzed CNN-based approaches for detecting and pinpointing synthetic image manipulations. Our Task A model underperformed (F1 = 0.0627), revealing a critical bias toward real images. This shows that binary classifiers struggle to generalize with mixed types of synthetic images. The segmentation part of Task B, however, was moderately successful at locating the localized changes, with an average F1 score of 0.357 and IoU of 0.3075. Our conclusion is that subsequent work should concentrate on addressing issues of dataset equilibrium, incorporating domain-specific data augmentation, and utilizing hybrid feature sets that integrate both spatial and frequency cues.

Acknowledgments

We thank the MediaEval 2025 organizers for providing the task datasets and evaluation framework.

Declaration on Generative AI

During this work, the author(s) used ChatGPT-4 and Grammarly for grammar and spelling checks. All content was reviewed and edited by the author(s), who take full responsibility for the publication.

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.

[2] 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, 2020.

[3] A. Mareen, et al., Tgif: A dataset for synthetic image manipulation detection, in: Proc. of the MediaEval Workshop, 2024.

[4] P. Zhou, X. Han, V. I. Morariu, L. S. Davis, Learning rich features for image manipulation detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2018).

[5] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, A. K. Roy-Chowdhury, Hybrid lstm and encoder-decoder architecture for image forgery localization, IEEE Transactions on Image Processing 28 (2019) 3286–3300.