An EfficientNet Framework : Methods and Results for Synthetic Image Detection and Manipulation Localization

Sakthi Mukesh Thanga Mariappan[1,*,†], Muthulakshmi Ramasamy[2,†] and Beulah Arul[1,†]

[1] *Rajalakshmi Engineering College, Chennai, Tamil Nadu, India*
[2] *Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India*

Abstract

This paper details the participation of the CodingSoft team in the MediaEval 2025 Synthetic Images Detection Challenge. We competed in two primary tasks: Task A, which focused on classifying images as either real or synthetically generated across both constrained and open runs, and Task B, which involved localizing manipulated regions within an image. Our approach utilized convolutional neural network (CNN) architectures, specifically leveraging EfficientNet, to build robust detection models. Our submissions yielded strong results in Task A, with our constrained run achieving an F1-score of 0.8485 and our open run achieving 0.8315. In Task B, our system for detecting manipulated images reached an average F1-score of 0.6523, while the localization model yielded an average Intersection over Union (IoU) of 0.326. This paper outlines our methodology, datasets, and a detailed comparative analysis of our results.

1. Introduction

The rapid advancement of generative AI [1] has led to a proliferation of highly realistic synthetic and manipulated images [2], posing significant challenges to digital information integrity. Detecting this content is crucial for combating misinformation and ensuring the authenticity of visual media. The MediaEval 2025 Synthetic Images Detection Challenge [3] provides a critical benchmark for evaluating methods designed to address this issue.

The challenge is divided into two sub-tasks:

- **Task A: Real vs. Synthetic Image Detection:** A classification task to determine if an image is entirely computer-generated or real.
- **Task B: Manipulated Region Localization:** A more granular task requiring both the detection of manipulation [4] and the generation of a pixel-level mask identifying the altered regions.

This paper presents the working notes of the CodingSoft-REC team, detailing our approach and performance in the challenge. We aim to contribute to the understanding of deep learning-based solutions for this complex problem.

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

[*] Corresponding author.

[†] These authors contributed equally.

sakthiimukesh@gmail.com (S. M. T. Mariappan); muthulakshmiiiir@gmail.com (M. Ramasamy); beulah.a@rajalakshmi.edu.in (B. Arul)

0009-0005-3044-6841 (S. M. T. Mariappan); 0009-0009-9911-5812 (M. Ramasamy); 0000-0002-3891-0806 (B. Arul)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Work

The detection of synthetic and manipulated media has evolved significantly, transitioning from traditional statistical methods to sophisticated deep learning approaches. Our work builds upon several key contributions in this domain. For synthetic image classification, Wang [5] demonstrated that standard ResNet architectures can effectively distinguish GAN-generated images by learning their intrinsic artifacts, particularly when trained with real-world augmentations like blur and JPEG compression. More recently, Yan [6] advanced diffusion-based image detection by analyzing latent denoising trajectories, revealing detectable signatures within the generative process itself.

In the domain of manipulation localization, several state-of-the-art methods have established strong baselines. Kwon [7] introduced a learning-based approach that leverages JPEG compression artifacts for both detection and localization of manipulated regions, demonstrating that compression-based features can be powerful discriminators. Building on multi-scale analysis, Dong [8] proposed MVSS-Net, which combines multi-view and multi-scale supervised networks to capture manipulation traces across different granularities. More recently, Li [9] presented UnionFormer at CVPR 2024, a unified transformer-based architecture that integrates multi-view representations for simultaneous detection and localization, achieving strong performance on benchmark datasets.

3. Dataset

The challenge utilized a diverse range of datasets to train and evaluate models under realistic conditions.

3.1. Task A: Real vs. Synthetic Image Detection

For the constrained run, participants were limited to two official training datasets:

- A benchmark set from Wang [5], containing images generated by older GAN models such as StyleGAN2 and BigGAN.
- A more recent dataset from Corvi [10], featuring 200,000 synthetic images from Latent Diffusion models and 200,000 real images from COCO and SUN.

For the open run,

- **CIFAKE Dataset** [11] which contains 60,000 real images from CIFAR-10 and 60,000 synthetic equivalents [12] generated with Stable Diffusion v1.4.

The validation and test sets each consisted of 10,000 images (5,000 real and 5,000 synthetic) collected from various online sources to simulate real-world diversity.

3.2. Task B: Manipulated Region Localization

The training data for this task was the TGIF dataset [13] (Mareen), which includes authentic images, ground truth masks, and manipulated versions created with different techniques:

- **Spliced (sp):** Manipulations generated using Adobe Photoshop (ps) and Stable Diffusion 2 (sd2).
- **Fully Regenerated (fr):** Images entirely regenerated by Stable Diffusion 2 (sd2) and Stable Diffusion XL (sdxl).

The validation data was sourced from COCO [14] and RAISE [15] and included original images alongside versions manipulated by seven different methods, such as brushnet, controlnet, and powerpaint.

4. Approach

For this task, we adopted a U-Net-like architecture with an EfficientNet backbone to perform both classification and segmentation. The model produces two outputs:

4.1. Task A: Real vs. Synthetic Image Detection

For constrained run, we used EfficientNet-B0, a lighter model pre-trained on ImageNet. Since the constrained run limited us to the official training datasets, we chose this architecture to prevent overfitting while maintaining strong generalization. The smaller parameter count of B0 was better suited to the relatively focused distribution of the official training data.

For open run, we scaled up to EfficientNet-B4, which has greater representational capacity. With access to the additional CIFAKE dataset, we hypothesized that the larger model could better learn the diverse range of generative signatures present across multiple data sources. However, our results showed that the increased model complexity did not translate to better performance, suggesting that model-data alignment is more critical than model size alone.

4.2. Task B: Manipulated Region Localization

This task required a model capable of both classification and segmentation. We adopted a U-Net-like architecture with an EfficientNet backbone. This design allowed the model to learn rich feature representations for detecting manipulation while preserving spatial information needed for accurate mask prediction. The loss function was a composite of:

- **Binary Classification Score:** A single scalar value indicating whether the image has been manipulated (values closer to 1 indicate manipulation, closer to 0 indicate pristine).
- **Pixel-Level Segmentation Mask:** A spatial map of the same dimensions as the input image, where each pixel value represents the probability that the corresponding region has been manipulated.

The model was trained with a composite loss function combining Binary Cross-Entropy Loss for the overall manipulation detection and Dice Loss for optimizing the pixel-wise mask prediction. This dual-output design allows the model to learn both global manipulation patterns and fine-grained localization simultaneously.

5. Resutls

Our models were evaluated on the test sets, and the results provide valuable insights into their performance.

5.1. Task A: Real vs. Synthetic Image Detection

We submitted runs for both the constrained and open categories, with the results offering a surprising insight into the role of external data. For constrained run, this model is trained only on the official datasets and achieved a very strong F1-score of 0.8485. For open run, the model

Table 1
Comparison of Constrained vs. Open Run Results for Task A

Metric	Constrained Run	Open Run
F1-Score	**0.8485**	**0.8315**
Accuracy	0.8592	0.8424
Precision	0.9181	0.8932
Recall	0.7888	0.7778
ROC AUC	0.9358	0.9185

trained with the CIFAKE dataset yielded slightly lower performance on the main evaluation metric. The bold numbers represent the strong scores achieved in the F1 score on both the runs.

5.1.1. Comparative Analysis:

Counterintuitively, our constrained run outperformed our open run across nearly all metrics. The F1-score was higher (0.8485 vs. 0.8315), as were accuracy and precision. This suggests that the external CIFAKE dataset, despite its large size, may have introduced a data distribution that did not align perfectly with the final test set, or that the simpler EfficientNet-B0 model was a more effective choice for the provided data. It highlights a critical insight: more data is not always better unless it is highly relevant to the target domain.

5.2. Task B: Manipulated Region Localization

The evaluation for Task B involved both detection and localization metrics. Our model's ability to detect whether an image was manipulated achieved an average F1-score of 0.6523 across the test set. For localization, measured by Intersection over Union (IoU), we achieved an average score of 0.326.

Performance varied significantly across different data sources and manipulation strategies. The model performed better on the RAISE dataset (average IoU of 0.452) compared to COCO (0.276), indicating potential dataset bias from our training on TGIF. More critically, we observed a clear performance gap between manipulation strategies: spliced manipulations, where only the inpainted region is modified, were generally easier to detect than fully regenerated images, where the entire image is synthetically reconstructed. This aligns with intuition, as fully regenerated images maintain greater visual coherence and consistency, making the boundaries of manipulation harder to identify. Among individual methods, our model struggled most with the controlnet technique but showed relatively stronger performance on hdpainter and the mixed manipulation combinations.

6. Conclusion

Our participation in the MediaEval 2025 Synthetic data challenge yielded valuable insights, particularly from the comparative results in Task A. Our EfficientNet-B0 model in the constrained run demonstrated excellent performance, surprisingly surpassing our open run model that was trained on more data. This finding underscores the importance of data quality and relevance over sheer quantity. For Task B, our model showed promise but also highlighted the significant challenge of creating a universally effective localization tool that can generalize across different manipulation methods.

Declaration on Generative AI

During the preparation of this work, the author used generative AI tools solely for grammar and spelling checks. All content—including approach, analysis, and discussion—has been prepared and critically reviewed by the author to ensure originality and clarity.

References

[1] 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, Springer, 2024, pp. 400–417.

[2] 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.

[3] 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.

[4] Z. Shi, X. Shen, H. Kang, Y. Lv, Image manipulation detection and localization based on the dual-domain convolutional neural networks, IEEE Access 6 (2018) 76437–76453.

[5] 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, pp. 8695–8704.

[6] A. Vasilcoiu, I. Najdenkoska, Z. Geradts, M. Worring, Latte: Latent trajectory embedding for diffusion-generated image detection, arXiv preprint arXiv:2507.03054 (2025).

[7] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, C. Kim, Learning jpeg compression artifacts for image manipulation detection and localization, International Journal of Computer Vision 130 (2022) 1875–1895. URL: http://dx.doi.org/10.1007/s11263-022-01617-5. doi:10.1007/s11263-022-01617-5.

[8] C. Dong, X. Chen, R. Hu, J. Cao, X. Li, Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 3539–3553.

[9] S. Li, W. Ma, J. Guo, S. Xu, B. Li, X. Zhang, Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12523–12533.

[10] 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.

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

[12] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).

[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, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[15] 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.