Multi-Modal Forensic Feature Fusion for Unified AI-Generated Content Detection and Manipulation Localization

J Bhuvana[1], Ramanan Mahendran[1] ,S Siddharth Chandrasekar[1] ,J Pragatheesh[1]

[1] *Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, India*

Abstract

This paper presents a unified deep learning pipeline addressing two critical challenges in multimedia forensics: Image-Level AIGC Detection and Manipulated Region Localization. For Task A, we propose a Hybrid CNN [2] and Vision Transformer (ViT) [4] architecture that fuses spatial, global, frequency-domain, and noise residual features to achieve robust binary classification against in-the-wild images and we utilize regularization (weight decay) and extensive data augmentation. For Task B, we implement a speed-optimized UNet [5] architecture with a ResNet34 [3] encoder for pixel-level semantic segmentation of altered regions. Our methodology demonstrates the feasibility of combining deep feature engineering with complex segmentation techniques to create a comprehensive digital integrity solution.

1. Introduction

The rapid advancement of generative AI models, such as Stable Diffusion [7] and Midjourney [14], has fundamentally altered the digital landscape by leading to an explosion of high-fidelity synthetic content. The overall approach is structured around two specific objectives.

The first, Task A (Real vs. Synthetic Image Detection), focuses on binary classification: determining whether an entire input image is authentic or synthetically generated. The second, Task B (Manipulated Region Localization), tackles a more granular problem: identifying if an image has been manipulated at the pixel level and producing a detailed probability mask that highlights the exact altered regions.

To achieve robust performance across both tasks, our pipeline leverages the complementary strengths of established Convolutional Neural Networks (CNNs) [2] and modern Vision Transformers (ViT) [4]. This strategic combination allows our system to extract deep forensic traces across multiple scales and domains—analyzing spatial structure, frequency-domain artifacts, and residual noise patterns.

2. Related work

2.1. Image Forgery Detection via Deep Features

Deep learning has accelerated the field of digital forensics by automating feature extraction. Architectures like MantraNet [6] and SRM-based CNNs [9] demonstrate that residual and noise layers are crucial. Recent research confirms that low-level forensic features are highly effective against GAN-based images [17], and can generalize across generative models [18].

The integration of Vision Transformers (ViT) [4] has further improved performance by capturing long-range dependencies that traditional CNNs [2] often miss. Our hybrid model in Task A builds upon this by fusing local (ResNet [3]), global (DeiT), and domain-specific (FFT [10]/Noise) features. Frequency-domain analysis captures subtle generative artifacts, offering a complementary view to spatial features.

2.2. Semantic Segmentation for Localization

Manipulated region localization is fundamentally a pixel-level segmentation problem. UNet [5] and its encoder-decoder variants (like those using ResNet [3] backbones) remain state-of-the-art. Traditional Image Forgery Localization (IFL) methods focused on detecting intrinsic artifacts, such as JPEG compression inconsistencies [19]. More recently, advanced deep learning models like MVSS-Net [20] and the unified learning approach UnionFormer [21] have achieved top-tier performance. The effectiveness of UNet [5] hinges on "skip connections," which transfer high-resolution feature maps from the contracting to the expanding path, preserving precise boundary information during decoding. Our implementation for Task B prioritizes speed optimization via input-size reduction and robust error handling, ensuring deployability and stability under competition constraints.

3. Approach

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

bhuvanaj@ssn.edu.in (J.Bhuvana); ramanan@ssn.edu.in (R.Mahendran); siddharth2312010@ssn.edu.in (S.Siddharth Chandrasekar); pragatheesh2312067@ssn.edu.in (J. Pragatheesh)

Our solution is divided into two distinct machine learning pipelines—one optimized for the classification objective of Task A, and another for the pixel-level segmentation required by Task B, with a strong emphasis on training stability and computational efficiency.

3.1. Task A: Hybrid AIGC Detection Methodology

The model architecture for Task A is a Hybrid CNN [2] and Vision Transformer (ViT) [4] designed to maximize forensic feature extraction. The network integrates core components including a ResNet50 [3] backbone to extract deep, local spatial features, and a DeiT Base (ViT) [4] encoder that captures global context and long-range dependencies via self-attention mechanisms. Custom feature-engineering streams using two linear layers process 2D-FFT frequency features [10] and Noise Residuals, targeting synthetic image artifacts. All four feature streams are concatenated before the final classification head.

For training stability and generalization, public benchmark datasets [13] and [14] were used, with the official 10k validation set for hyperparameter tuning. Data augmentations included RandomHorizontalFlip, RandomRotation (15°), and ColorJitter (brightness=0.2, contrast=0.2). Optimization used the Adam optimizer with L2 regularization (Weight Decay) to prevent overfitting. The model was trained for 8 epochs using a batch size of 16.

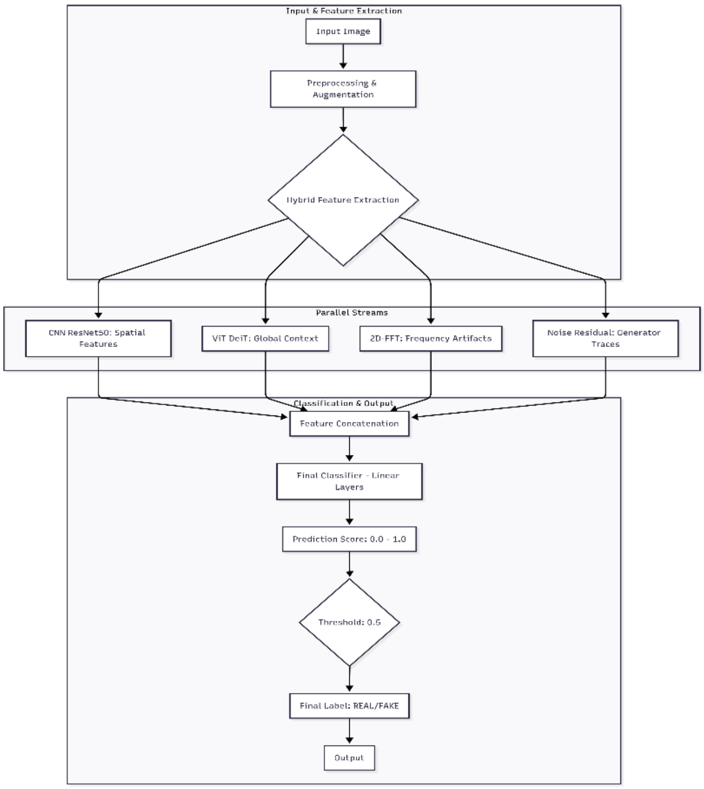

Figure 1: Pipeline for Hybrid AIGC Detection (Task A) showing multi-modal feature fusion for binary classification.

3.2. Task B: Region Localization Methodology

Task B utilized a specialized segmentation architecture to achieve pixel-level localization. The chosen architecture was a UNet [5] variant employing a ResNet34 [3] encoder backbone. This configuration is preferred in segmentation tasks for its ability to maintain fine spatial resolution through *skip connections* between the encoder and decoder stages, enabling precise boundary reconstruction and accurate pixel-level localization. Training was conducted using the TGIF dataset, which includes a diverse range of spliced and fully regenerated images paired with binary ground-truth masks. The official validation set, featuring COCO [15] and RAISE [16] images altered by methods such as *BrushNet* and *ControlNet*, was reserved strictly for evaluation. Due to competition-imposed hardware limits, aggressive computational optimization was necessary. A crucial design choice for speed optimization involved reducing input and mask resolutions, ensuring a feasible training time of approximately 15–20 minutes per epoch without compromising model convergence. The segmentation model was optimized using BCEWithLogitsLoss, chosen for its numerical stability and robustness in binary pixel-classification scenarios.

The image-level manipulation prediction (reported in Table 2) was derived from the mean probability of the final segmentation mask. If the average predicted probability of all pixels exceeded a global threshold of 0.5, the image was classified as manipulated; otherwise, it was considered authentic. This approach provided a consistent mechanism for converting fine-grained pixel predictions into global binary decisions, aligning the segmentation output with Task B's evaluation metrics.

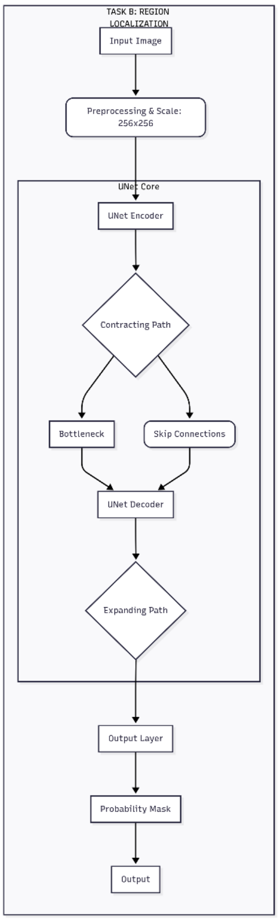

Figure 1
Conceptual Flow for Manipulated Region Localization (Task B) via UNet [5] based pixel-level semantic segmentation.

4. Results and Analysis

4.1. For Task B (Manipulation Localization), the quantitative results are summarized in Table 3, detailing detection metrics (Balanced Accuracy, AUC, F1), and Table 4, detailing localization metrics (F1, IoU).

Table 1
Result score for the challenge run 1

	Pred Real	Pred Fake	Total
REAL	3559	1441	5000
FAKE	545	4454	4999
TOTAL	4104	5895	9999

Table 2
Metric score for the challenge run 1

Metric	Score
Accuracy	0.8014
F1-Score	0.8177

Metric	Score
ROC-AUC	0.9032

4.2. For Task B (Manipulation Localization), the quantitative results are summarized in Table 3, detailing detection metrics (Balanced Accuracy, AUC, F1), and Table 4, detailing localization metrics (F1, IoU).

Table 3
Result score for the challenge run 2 – detection

Source	Method	Balanced Accuracy	AUC	F1	AP
AVG	AVG	0.5000	0.4941	0.6917	0.5158
raise	AVG	0.5000	0.4817	0.8521	0.7345
coco	AVG	0.5000	0.5017	0.6667	0.4877
openimages	AVG	0.5000	0.5059	0.6673	0.5005

Table 4
Result score for the challenge run 2 - localization

Folder	Method	F1_best	F1_th	IoU
raise	AVG	0.443004	0.4382104609	0.31170464
coco	AVG	0.28694466	0.2795308622	0.19363518
openimages	AVG	0.3299671	0.3211907837	0.23489884
AVG	ALL	0.33617887	0.3284522725	0.23566236

4.3. Summary and Insights

Our results for Task A showed strong overall performance with an accuracy of 0.801, F1-score of 0.8177, and excellent discriminative power, indicating the hybrid CNN [2] /ViT [4] architecture effectively fused multi-modal features to distinguish explicit generative artifacts. For Task B, binary detection exhibited high recall but low precision, reflecting over-sensitivity. Pixel-level localization (IoU) [12] was significantly impacted by the resolution constraint, leading to limited precision, particularly for subtle manipulations or specific datasets like. While the UNet [5] was efficient, this speed-accuracy trade-off highlights that fine-grained localization remains challenging, underscoring the need for future exploration into adaptive resolution techniques to balance computational efficiency with granular accuracy in real-world scenarios.

5. Conclusions

We successfully developed and implemented comprehensive deep learning pipelines for both image-level AIGC detection (Task A) and pixel-level manipulation localization (Task B). The Task A hybrid model effectively fused CNN and Transformer feature to enhance detection robustness, while the Task B pipeline achieved efficient segmentation with minimal computational overhead. Future work will explore multiscale feature fusion and hierarchical training to recover fine-grained accuracy lost due to resolution reduction, ensuring improved localization without exceeding competition time limits.

Acknowledgments

We thank the organizers of the MediaEval 2025 task for providing the benchmark and evaluation framework, and acknowledge the guidance of Dr. J. Bhuvana and the computational support from the Department of Computer Science and Engineering, SSN College of Engineering.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Papadopoulou, O., Schinas, M., Corvi, R., Karageorgiou, D., Koutlis, C., Guillaro, F., Gavves, E., Mareen, H., Verdoliva, L., and Papadopoulos, S. (2025). Synthetic Images at MediaEval 2025: Advancing Detection of Generative AI in Real-World Online Images. Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 25–26 October 2025.

[2] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

[3] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[4] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning (ICML), Vol. 139, 10347–10357.

[5] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234–241.

[6] Wu, Y., Abd-Almageed, W., and Natarajan, P. (2019). Mantra-Net: Manipulation Tracing Network for Detection and Localization of Diverse Image Forgeries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 910–919.

[7] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.

[8] Zhang, L., Rao, A., and Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3836–3847.

[9] Bayar, B., and Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. In Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec), 5–10.

[10] Fan, C.-C., Shuai, H.-H., and Chen, W.-C. (2020). FDF-Net: A Fingerprint Detection and Fusion Network for Image Forgery Detection. IEEE Signal Processing Letters, 27, 1820–1824.

[11] Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

[12] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2), 303–338.

[13] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8695–8704.

[14] Corvi, R., Cozzolino, D., Poggi, G., and Verdoliva, L. (2023). On the Detection of Synthetic Images Generated by Diffusion Models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5.

[15] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In European Conference on Computer Vision (ECCV), 740–755.

[16] Dang-Nguyen, D.-T., Pasquini, C., Conotter, V., and Boato, G. (2015). RAISE: A Raw Images Dataset for Digital Image Forensics. In Proceedings of the 6th ACM Multimedia Systems Conference (MMSys), 219–224.

[17] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[18] Ojha, U., Zhang, W., and Wang, H. (2023). Towards universal fake image detectors that generalize across generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[19] Kwon, H., Kim, K., and Kim, Y. (2022). Learning JPEG compression artifacts for image manipulation detection and localization. International Journal of Computer Vision (IJCV).

[20] Dong, X., Wang, Q., Liu, Q., Li, W., and Zhang, J. (2022). MVSS-Net: Multi-view Multi-scale Supervised Networks for Image Manipulation Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).