A resolution-agnostic three-stage framework for Image Forgery Detection and Localization

Minh-Hoang Le[1,3,†], Minh-Khoa Le-Phan[1,3,†], Khanh-Ngoc Vu-Nguyen[2,3], Minh-Triet Tran[1,3,*] and Trong-Le Do[1,3,*]

[1] *University of Science - VNU-HCM, Ho Chi Minh City, Vietnam*
[2] *High School for the Gifted - VNU-HCM, Ho Chi Minh City, Vietnam*
[3] *Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract

With the development of generative models and text-guided inpainting methods, current SID and IFL methods are gradually becoming outdated. In this paper, we propose a resolution-agnostic three-stage framework for the MediaEval 2025 SynthIM challenge. Our pipeline first detects FR versus OR/SP, then localizes manipulated regions, and finally classifies OR versus SP using the output of the localization stage. Our method works well and gives strong results on both the classification task and the localization task.

1. Introduction

Along with the progress of generative models, text-guided inpainting techniques are becoming more realistic and more semantically aligned to the texture. For that reason, the need for stronger and more general models to detect manipulated images is also rising. Current research is divided into two subtasks: Synthetic Image Detection (SID), which only decides whether an image is real or fully generated without any convincing reason, and Image Forgery Localization (IFL), which points out manipulated regions in partially modified images. However, current IFL methods ignore cases where an image is fully regenerated instead of being blended back into the original, which leads to missing reliable evidence of manipulation.

To address these limits, MediaEval 2025 introduced the SynthIM challenge [1]. The organizers challenge participants with two subtasks: first, build a model to classify whether an image is real or manipulated, and second, localize the manipulated regions.

In this paper, we present a resolution-agnostic three-stage framework to classify and localize manipulated regions for three types: original, spliced, and fully regenerated. Our framework emphasizes the ability to infer on images of any resolution and its flexible pipeline.

2. Related Work

Current SID methods tend to detect arbitrary resolution images with the patch-based approach. SPAI [2] tries to model the distribution of real images with low-resolution training images and infer any-resolution images via inferring patches. TextureCrop [3] leverages the sliding window technique to divide an image into patches, ranks them, chooses the K top patches, and aggregates the results of all K patches.

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

[*] Corresponding authors.

[†] Both authors contributed equally to this research as co-first authors.

lmhoang22@apcs.fitus.edu.vn (M. Le); lpmkhoa22@apcs.fitus.edu.vn (M. Le-Phan); student230319@ptnk.edu.vn (K. Vu-Nguyen); tmtriet@fit.hcmus.edu.vn (M. Tran); dtle@selab.hcmus.edu.vn (T. Do)

0009-0005-1501-8080 (M. Le); 0009-0005-9707-4026 (M. Le-Phan); 0009-0005-8210-1829 (K. Vu-Nguyen); 0000-0003-3046-3041 (M. Tran); 0000-0002-2906-0360 (T. Do)

CEUR Workshop Proceedings (CEUR-WS.org)

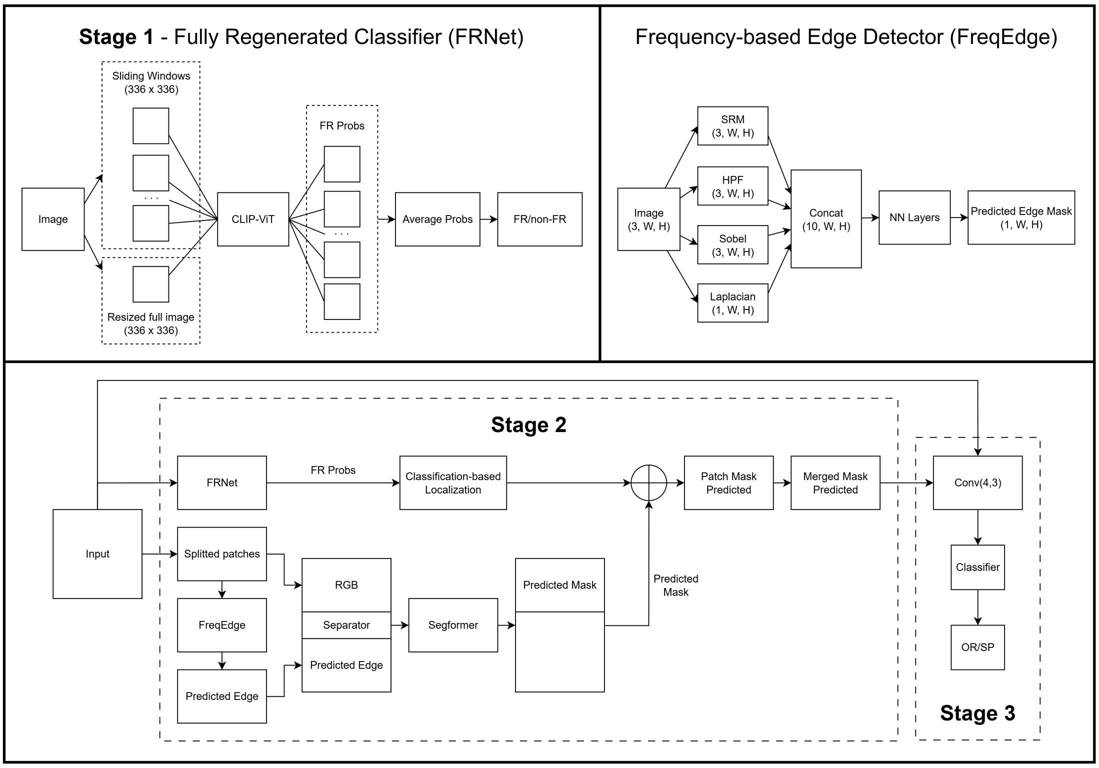

Figure 1: Overall pipeline of our proposed framework.

Some IFL methods try using frequency-based and edge-based cues to improve the results. HiFi [4] proposes a frequency block in parallel with the color block to extract frequency features. DiffForensics [5] introduces an edge loss to guide the proposed model to learn more efficiently.

3. Approach

We design a three-stage framework for the **Manipulated Region Localization** track of the MediaEval 2025 SynthIM challenge [1], where the goal is to localize synthetic regions in images that may be degraded by resizing, cropping, or compression. Following the dataset protocol, we distinguish between **Original (OR)**, **Spliced (SP)**, and **Fully Regenerated (FR)** images [6, 7]. Since FR images contain global generative artifacts, they are handled as a classification problem, while SP images contain localized edits that provide explicit masks for evaluation. Our system therefore combines (i) two classification stages for detection, and (ii) one dual-branch localization stage for SP detection and localization.

3.1. Stage 1: FR versus OR/SP Classification

We first perform a coarse FR vs. OR/SP classification before localization. SP–OR separation is challenging due to small edited regions and downscaling of ultra–high-resolution images (up to 10K), which weakens subtle cues. In contrast, FR images differ more clearly and often show semantic artifacts. Thus, we train an FR-vs-OR/SP classifier to guide localization (Section 3.2.2).

We first train **FRNet** (Figure 1) by applying LoRA [8] fine-tuning on CLIP-ViT [9] to distinguish Fully Regenerated (FR) from non-FR content. During training, inputs are randomly cropped to 336×336 with random horizontal flipping for augmentation. At test time, we employ a sliding-window strategy with window size 336×336, and also include the whole image resized to 336×336 as an additional window. Each window is processed by the trained CLIP-ViT to output a probability, and the final FR probability is obtained by averaging across all windows.

This model achieves 97.5% accuracy on the validation set and serves as the first gate of our pipeline. With the trained FRNet, classification on the test set proceeds as follows. Each test image is assigned a global FR probability p_{FR}. If p_{FR} exceeds a validation-tuned threshold, the image is labeled FR; the probability p_{FR} is recorded in the `scores.csv` file.

3.2. Stage 2: Localization for OR/SP Images

For images not classified as FR, we predict pixel-level masks using a **dual-branch localization model** (Figure 1) that integrates boundary and region cues. Although this model is primarily designed for OR/SP images, we also apply it to FR images to generate their localization masks. In other words, we currently do not employ a dedicated localization model for FR images, but reuse the same model used for OR/SP images.

3.2.1. Boundary-Aware Localization (BAL)

The BAL branch combines frequency-based boundary priors with semantic segmentation.

Edge detector. We build a Frequency-based Edge Detector (FreqEdge) (Figure 1) to reveal blending traits of spliced images. This detector operates on residual filters and fixed Sobel/Laplacian gradients, producing a frequency-based edge prior. The residual and gradient responses are then fed into lightweight neural layers, which fuse these cues and output an edge probability map. In this way, FreqEdge highlights transition bands around manipulated regions while remaining robust to scene semantics.

SegFormer integration. The soft edge probability map is stacked with the RGB image and injected into a SegFormer decoder to produce a region-level probability map. This integration sharpens boundaries and stabilizes contours during mask prediction.

3.2.2. Classification-Based Localization (CBL)

To complement BAL, we reuse the FR classifier in a sliding-window mode to produce dense probability heatmaps of synthetic evidence. Overlapping windows are fused by Hann-weighted blending, yielding smooth region maps even when no boundary is visible. This branch improves robustness on high-resolution images and uniform patches.

3.2.3. Fusion

The two maps are combined by pixel-wise maximum:

$$\mathbf{P}_{\text{fuse}}(u,v) = \max\big(\mathbf{P}_{\mathrm{BAL}}(u,v),\ \mathbf{P}_{\mathrm{CBL}}(u,v)\big),$$

to produce the final mask $\hat{\mathbf{Y}}$, which is saved for the submission. This ensures that strong evidence from either branch is retained.

3.3. Stage 3: OR versus SP Classification

In addition to the pixel-level mask, we assign a global manipulation probability for the `scores.csv` submission. To this end, we concatenate the fused localization map with the RGB image (3+1 channels), project back to 3 channels via a lightweight adapter, and feed the result into three complementary classifiers (EfficientNet-B4 [10], CvT-w24 [11], CLIP-ViT [9]). Feature vectors are concatenated and classified into OR or SP, and their corresponding scores are saved to the CSV file.

In summary, our approach exploits complementary cues: FR classification handles global generative artifacts, while the dual-branch localization model (BAL + CBL) provides fine-grained detection and localization for OR/SP images. This design balances robustness on high-resolution, degraded inputs with accurate region-level localization.

Table 1
Stage 1: Classification performance of FRNet using two CLIP-ViT backbones on the validation set. We compare models with input resolutions 224×224 and 336×336 for the task of distinguishing Fully Regenerated (FR) from non-FR (OR/SP) images.

Model	Accuracy
1) CLIP-ViT 224	0.952
2) CLIP-ViT 336	**0.975**

Table 2
Stage 3: Classification performance for distinguishing Original (OR) from Spliced (SP) images on the validation set. We evaluate three individual backbones (EfficientNet-B4, CvT-w24, and CLIP-ViT 336) as well as their ensemble. The ensemble achieves the highest accuracy.

Model	Accuracy
1) EffB4	0.865
2) CvT-w24	0.882
3) CLIP-ViT 336	0.895
4) Ensemble: (1) + (2) + (3)	**0.916**

4. Results and Analysis

Table 1 and Table 2 present our results in the classification subtask. Our method achieves 81.3% F1 score and 74.4% balanced accuracy in the official test set.

Table 3
Stage 2: Manipulated localization performance. **IoU (SP validation)** is calculated by using SP images in the validation set as the ground truth. **IoU (full test)** is reported by the organizer on the manipulated test set (FR+SP). We present four settings: 1) Segformer with RGB as input; 2) Segformer with a 4-channel RGB(3)+Edge(1) input; 3) (2) plus the Heatmap branch; 4) Use (1) for images classified in Stage 1 as FR and (3) otherwise.

Model	IoU (SP validation)	IoU (full test)
1) Segformer (RGB)	0.298	0.488
2) Segformer (RGB + Edge)	0.440	0.459
3) (2) + Heatmap	**0.602**	0.467
4) Ensemble: **FR** (1) + **SP OR** (3)	**0.602**	**0.515**

Table 3 summarizes our results in the localization subtask. We note an interesting pattern: on the SP validation set, setting (1) underperforms settings (2) and (3), yet on the full test set, it achieves a higher IoU than both. This appears inconsistent with our SP validation visualizations, where (1) is clearly worse. The discrepancy stems from the presence of FR images in the test set. To address this, we introduce setting (4), an ensemble that uses setting (1) for images predicted as FR in Stage 1 and setting (3) for the remaining images, which yields the best test performance.

5. Discussion and Outlook

We introduced a simple, resolution-agnostic three-stage system for the SynthIM task. Stage 1 detects FR images, Stage 2 localizes manipulated regions, and Stage 3 uses the localization map to decide OR versus SP. Using both edge cues and heatmap evidence helped make the masks more robust across different image sizes. We also found that choosing different localization models for predicted FR and SP images improves test performance.

Acknowledgments

This research is funded by Vietnam National University, Ho Chi Minh City (VNU-HCM) under grant number DS.C2025-18-13.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. 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.

[2] D. Karageorgiou, S. Papadopoulos, I. Kompatsiaris, E. Gavves, Any-resolution ai-generated image detection by spectral learning, 2025, pp. 18706–18717. doi:10.1109/CVPR52734.2025.01743.

[3] D. Konstantinidou, C. Koutlis, S. Papadopoulos, Texturecrop: Enhancing synthetic image detection through texture-based cropping, 2025, pp. 1369–1378. doi:10.1109/WACVW65960.2025.00160.

[4] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, X. Liu, Hierarchical fine-grained image forgery detection and localization, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3155–3165. doi:10.1109/CVPR52729.2023.00308.

[5] Z. Yu, J. Ni, Y. Lin, H. Deng, B. Li, Diffforensics: Leveraging diffusion prior to image forgery detection and localization, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 12765–12774. doi:10.1109/CVPR52733.2024.01213.

[6] H. Mareen, D. Karageorgiou, G. V. Wallendael, P. Lambert, S. Papadopoulos, Tgif: Text-guided inpainting forgery dataset, in: 2024 IEEE International Workshop on Information Forensics and Security (WIFS), 2024, pp. 1–6. doi:10.1109/WIFS61860.2024.10810690.

[7] P. Giakoumoglou, D. Karageorgiou, S. Papadopoulos, P. C. Petrantonakis, Sagi: Semantically aligned and uncertainty guided ai image inpainting, 2025. URL: https://arxiv.org/abs/2502.06593. arXiv:2502.06593.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

[10] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[11] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: Introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 22–31.