A CLIP-based Approach for Synthetic Image Detection under Distribution Shift

Qiushi Li[1,*], Andrea Ciamarra[2], Roberto Caldelli[2,3] and Stefano Berretti[1]

[1]*Media Integration and Communication Center (MICC), University of Florence, Italy*
[2]*CNIT, Florence, Italy*
[3]*Mercatorum University, Rome, Italy*

Abstract

For the MediaEval 2025 SynthIm Challenge Task A (Synthetic Image Detection), we identified a critical distribution shift between the provided training and test data, making the former a poor representative for the target domain. Our approach directly addresses this challenge by incorporating the more stylistically consistent validation set into our training data and leveraging a frozen CLIP ViT-L/14 as a robust feature extractor. Our main insight is that under such domain shift conditions, the generalizable representations from a large pre-trained model significantly outperform a traditional CNN fine-tuned on the mismatched dataset, proving to be a more effective and reliable strategy.

1. Introduction

The rapid advancement of generative models, including Generative Adversarial Networks (GANs), diffusion models, and large multimodal models, has enabled the creation of highly realistic synthetic media. While these technologies offer immense creative potential, they also pose significant challenges related to misinformation, disinformation, and malicious use. Consequently, the development of robust and reliable methods for detecting synthetic images has become a critical area of research in digital forensics and media security.

The MediaEval 2025 Synthetic Images Detection Challenge (SynthIm) [1] provides a timely and relevant benchmark for evaluating the state-of-the-art in synthetic image detection. Task A specifically focuses on the binary classification of images as either real or fake (synthetic). A key challenge in this field is the generalization capability of detectors. Generative models evolve rapidly, and a detector trained on images from one set of models may fail to perform well on images generated by newer, unseen architectures.

In our preliminary analysis of the task dataset, we identified a critical issue: a pronounced domain shift between the officially provided training and test sets. The visual characteristics, such as style and potential artifacts, present in the training samples appeared inconsistent with those in the test samples, which a human observer could easily perceive. In contrast, the validation set exhibited a visual distribution remarkably similar to that of the test set. This observation forms the cornerstone of our work. We hypothesize that the official training set may offer little benefit for detection in the target domain, and could even be detrimental, as it may cause a model to overfit to irrelevant features not present in the test set.

In this work, we first identify and analyze a significant distribution gap between the training and test sets provided in the SynthIm Task A dataset. To address this issue, we incorporate the validation set into the training process to better approximate the test distribution. We then

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

[*] Corresponding author.

qiushi.li@unifi.it (Q. Li)

CEUR Workshop Proceedings (CEUR-WS.org)

conduct a comparative study between a traditional, fine-tuned CNN detector (ResNet-50 [2]) and a modern approach that leverages a large-scale pre-trained model (CLIP ViT-L/14 [3, 4]). Our experiments demonstrate that using a frozen CLIP model as a feature extractor provides superior performance, validating the hypothesis that its rich, general-purpose features are more robust to domain shift than learning from a mismatched training set.

2. Related Work

The field of synthetic image detection has rapidly progressed from approaches that target generator-specific artifacts to those striving for universal generalization. Early studies, such as Wang et al. [5], demonstrated that images produced by early CNN-based generators exhibited consistent spectral artifacts, enabling reliable detection. However, detectors that depend on such specific artifacts often struggle to generalize to unseen generative architectures. To address this limitation, more recent research—exemplified by Ojha et al. [6]—has shifted toward developing universal detectors. This modern paradigm emphasizes leveraging large-scale pre-trained models like CLIP [3, 4], which learn robust and semantically meaningful visual representations rather than overfitting to low-level, model-specific cues. Our work follows this line of thought, employing a frozen CLIP model and the validation set to effectively handle the severe distribution shifts present in the dataset used for this task.

3. Results and Analysis

We submitted nine runs in total, including two constrained runs and seven open runs, with the following configurations:

- Run #1: Following the approach of Wang et al. [5], we trained a ResNet-based model using the official training set. Data augmentation included blur and JPEG compression with a probability of 0.5.
- Run #2: Identical to Run #1, except that the augmentation probability was reduced to 0.1.
- Run #3: Starting from the model trained in Run #1, we further fine-tuned it using 2,000 real and 2,000 fake samples randomly selected from the validation set provided by the organizers.
- Run #4: Similar to Run #3, but fine-tuned from the model in Run #2 instead of Run #1.
- Run #5: Inspired by Ojha et al. [6], we adopted a CLIP-based deepfake detection method. Specifically, we extracted image embeddings using a frozen CLIP model and added a trainable linear layer as the classifier. The model was first pre-trained on the official training set and then fine-tuned on the aforementioned 2,000 pairs of validation samples.
- Run #6: Similar to Run #5, except that the final linear layer was randomly initialized and trained directly on the 2,000 validation pairs—-i.e., the official training set was not used.
- Run #7: Based on the same setting as Run #6, but we increased the fine-tuning set to 4,000 validation pairs.
- Run #8: Note that in all previous runs, the prediction threshold was set to 0.5. In this run, we applied a majority voting strategy across the predictions of Runs #3, #4, #6, and #7 to enhance generalization.
- Run #9: Extending Run #8, we employed a weighted majority voting strategy, where the weights corresponded to the F1 scores of Runs #3, #4, #6, and #7 on the test set.

Table 1
Performance of various runs. The first set of metrics uses a default threshold, while the second set is optimized for the best F1 score by adjusting the threshold. Performance is evaluated using metrics where Accuracy, Precision, Recall, and F1-score are abbreviated as Acc, P, R, and F1, respectively.

	Default Threshold (0.5)				Best Global Threshold by F1				
Run Name	Acc	P	R	F1	thr	F1	P	R	Acc
Run #1	0.4926	0.3844	0.0246	0.0462	0	0.6667	0.5000	1.0000	0.5000
Run #2	0.4994	0.4118	0.0028	0.0056	0	0.6667	0.5000	1.0000	0.5000
Run #3	0.8449	0.9023	0.7736	0.8330	0.146	0.8530	0.8697	0.8370	0.8558
Run #4	0.8461	0.8905	0.7892	0.8368	0.083	0.8504	0.8285	0.8734	0.8463
Run #5	0.7406	0.8126	0.6254	0.7068	0.074	0.7540	0.6990	0.8184	0.7330
Run #6	0.8844	0.9206	0.8414	0.8792	0.398	0.8868	0.8686	0.9058	0.8844
Run #7	0.8894	0.9175	**0.8558**	0.8856	0.384	**0.8945**	0.8755	**0.9144**	0.8922
Run #8	0.8825	0.9281	0.8292	0.8759	0.351	0.8902	0.8897	0.8906	0.8901
Run #9	**0.8923**	**0.9329**	0.8454	**0.8870**	0.403	0.8943	**0.8960**	0.8926	**0.8945**

Table 1 summarizes the performance of all runs. It is evident that Run #7 and Run #9 achieved the best overall results. Under a fixed prediction threshold of 0.5, their F1 scores reached 0.8856 and 0.887, respectively. When evaluated at the best global threshold by F1, their F1 scores were 0.8945 and 0.8943, respectively.

Specifically, Runs #1 and #2, which are CNN-based methods, yielded almost random, ineffective predictions, indicating that training solely on the official training set is not enough for this task. Employing stronger backbones, such as CLIP-ViT in [6], still struggles to generalize to real-world scenarios when training on a narrow domain-specific dataset, achieving nearly random guess performance, as confirmed in the results in Table 1 of [1]. Therefore, CNN and CLIP-ViT based methods trained on their official datasets could not be sufficient to generalize well due to the severe distribution gap between training and test sets. From a perceptual perspective, we closely examined images in the training, validation, and test sets. It is visually apparent that the validation and test images share highly similar styles, and according to the challenge description, both were collected from the wild. Therefore, involving validation samples in training can be beneficial for improving test-time performance.

By comparing Run #5 and Run #6, we infer that training detectors on mismatched datasets can severely harm performance on real-world, social-media-derived test images. In contrast, incorporating more semantically relevant data—-such as the additional validation samples in Run #7-—significantly improves detection accuracy. Although ensemble methods are generally effective for enhancing performance and robustness, when the validation and test sets are highly similar in distribution, the performance gap becomes marginal—as reflected in the comparable results of Run #7 and Run #9.

4. Discussion and Outlook

In our work for the MediaEval 2025 SynthIm task, we successfully addressed the core challenge of a significant distribution shift between the training and test data. Our approach achieved desirable results by incorporating the more stylistically representative validation set into the training process and by utilizing a frozen CLIP ViT-L/14 model as a powerful feature extractor. The experiments clearly demonstrate that in a domain shift scenario, relying on the general and robust features of a large-scale pre-trained model is a more effective and reliable strategy than end-to-end fine-tuning on a mismatched dataset. This work also reveals that in the field of synthetic image detection, a deep understanding of the data itself is as important as the selection of an advanced model architecture.

Acknowledgment

This work was partially supported by the following projects: AI4Debunk (GA n. 101135757) funded by EU Horizon Europe Programme, SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU and FOSTERER funded by the Italian MUR PRIN-2022.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: check and improve grammar, spelling, and phrasing. After using this tool, 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] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[4] 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, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.

[5] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot… for now, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8692–8701. doi:10.1109/CVPR42600.2020.00872.

[6] U. Ojha, Y. Li, Y. J. Lee, Towards universal fake image detectors that generalize across generative models, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24480–24489. doi:10.1109/CVPR52729.2023.02345.