Real Versus Synthetic Classification using ResNet

Hafiz Muhammad Owais Raza[1,*,1], Muhammad Atif Tahir[1,1] and
Rizwan Ahmed Khan[1,1]

[1] *School of Mathematics and Computer Science(SMCS), IBA, Karachi, Pakistan*

Abstract
This paper presents our approach for the MediaEval 2025 “Real vs. Synthetic Image Classification” task, focusing on the use of deep convolutional networks for discriminating AI-generated images from authentic photographs. We employ ResNet-based architectures to explore the effectiveness of residual learning for robust feature extraction and generalization. Our experiments investigate two complementary training regimes designed to evaluate how dataset diversity and model capacity influence detection reliability. Through systematic training, augmentation, and inference strategies, we assess the ability of ResNet models to identify subtle artifacts introduced by generative models. The results highlight that deeper residual architectures combined with diverse training data significantly enhance detection accuracy and robustness against unseen synthetic sources. We conclude with insights on improving transferability and cross-generator generalization in real–synthetic image classification.

1. Introduction

This paper presents our contribution to the MediaEval 2025 “Real vs. Synthetic Image Classification” task [1], which focuses on distinguishing AI-generated images from authentic photographs. Our approach leverages convolutional neural networks (CNNs), particularly ResNet-based architectures, to detect subtle artifacts and distributional inconsistencies introduced by generative models. We evaluate two complementary strategies: a lightweight ResNet-18 for efficient feature extraction on domain specific data and a deeper ResNet-50 trained on a more diverse dataset for enhanced generalization. The comparative analysis highlights how model depth and data variety influence performance in cross-generator synthetic image detection.

2. Related Work

ResNet and its variants remain popular backbones for fake image detectors due to their strong feature learning [2]. Many methods fine-tune off-the-shelf networks (ResNet-50/101, etc.) or modify them to better capture artifacts. A common approach is to leverage the multi-scale feature hierarchy of ResNets. For example, Ju et al. [3] divided a ResNet-50 model into multiple layer groups to separately extract low-level vs high-level features, then used a patch selection module to identify “high-energy” regions before fusing local and global features for the final classification. This helps the network zoom in on fine detail artifacts (e.g. pixel inconsistencies) while also considering the overall image context. Similarly, Zhang et al. [4] proposed a three-branch ResNet-based architecture that alternates training between the main backbone and auxiliary branches, improving the model’s capacity to capture subtle cues across different feature scales. These works show that ResNet backbones can be enhanced with custom modules (patch-wise analysis, multi-branch fusion, etc.) to better detect AI-generated traces.

Ojha et al. [5] analyzed this failure mode in detail: they found that a standard ResNet classifier trained to distinguish real vs StyleGAN images learns highly specific low-level “fingerprints”

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

hm.owaisraza.29637@khi.iba.edu.pk (H. M. O. Raza); atiftahir@iba.edu.pk (M. A. Tahir); rizwankhan@iba.edu.pk (R. A. Khan)

CEUR Workshop Proceedings (CEUR-WS.org)

of the StyleGAN. When presented with an image from a completely different generator (e.g. DALL · E 2 or Stable Diffusion), the classifier often cannot recognize it as fake because the image lacks those specific fingerprints.

In summary, generalization remains the core challenge in real vs fake image classification. Efforts to improve it include: increasing training diversity (thousands of generators, many real sources), using stronger pre-trained feature extractors [6], crafting better augmentations and loss functions to avoid overfitting, and applying strategies such as those explored in MediaEval 2025 [1].

3. Approach

3.1. Architectures

We adopt residual convolutional networks as the foundation for real–synthetic image discrimination. Two model configurations were investigated: (1) a lightweight **ResNet-18** used for efficiency over a domain restricted data , and (2) a deeper **ResNet-50** over a diverse dataset to explore the effect of model capacity and data diversity on generalization. Additionally, a small-scale experiment with a Vision Transformer (ViT-B/16) was conducted to test the ability of self-attention to capture long-range spatial dependencies, though results remained qualitative due to compute constraints.

3.2. Datasets and Experimental Setup

All experiments were conducted using the official datasets provided by the MediaEval 2025 "Real vs. Synthetic Image Classification" task **mediaeval_overview**. In accordance with the task guidelines, both configurations presented in this work are based primarily on the official training data and therefore correspond to two distinct *constrained runs* that differ in their data scope and augmentation strategies.

Configuration 1: Constrained run setup. The first configuration was trained exclusively on a selected subset of the official labeled corpus. This subset comprised approximately **700,000 images** evenly distributed across **20 semantic categories**, containing an equal balance of real and synthetic examples. The purpose of this setup was to examine model performance when trained under limited domain diversity and controlled class representation. The corresponding validation set was used strictly for performance evaluation and was not incorporated into the training process.

Configuration 2: Open run setup. The second configuration also relied primarily on the official MediaEval training data but was designed to explore enhanced generalization through data diversity. In this setup, the base dataset was enriched with extensive image augmentations—including random cropping, horizontal flipping, color jittering, Gaussian blur, and slight rotation—to simulate broader appearance variability. To further approximate open-domain conditions, a small supplementary set of synthetic images was generated using publicly available AI tools such as *Chatgpt*. Moreover, a randomly selected subset of approximately **2,000 images** from the official validation set was employed for limited fine-tuning and hyper-parameter calibration, while the remaining validation data were reserved exclusively for evaluation.

Test data. The official unlabeled test set, consisting of **10,000 images**, was used only during inference for final prediction generation and submission formatting. No test samples were used in any training or tuning stage.

All images were resized to 224 × 224, normalized using ImageNet statistics, and augmented with random horizontal flips, light color jitter, and Gaussian blur. Aggressive augmentations (Mixup, CutMix) were tested but later discarded due to reduced calibration stability.

3.3. Training Configuration

All models were implemented in PyTorch using pretrained ImageNet weights for initialization. Training employed the AdamW optimizer with a learning rate of 1×10^{-4}, cosine annealing scheduler, and weight decay of 1×10^{-2}. The batch size was 32 and each model was trained for up to 25 epochs with early stopping based on validation ROC-AUC. Five-fold stratified cross-validation was conducted on the labeled data to assess stability and variance. Each experiment was conducted with a fixed random seed (42), ensuring consistent data splits and model initialization. The final model for each regime was retrained on all labeled images before test-time inference.

Hardware. Experiments were conducted on a macOS environment equipped with a Radeon Pro 5600M GPU using Apple's Metal Performance Shaders (MPS) backend. CPU fallback was used for preprocessing and small validation batches.

4. Results and Analysis

4.1. Constrained Run (ResNet-18)

The ResNet-18 model demonstrated rapid convergence and achieved high training accuracy on the initial dataset. However, its validation performance declined noticeably when evaluated on images representing visual domains not present in the training data. This behavior indicates that the network tended to overfit domain-specific visual characteristics rather than capturing generalizable cues that distinguish real from synthetic imagery—such as subtle inconsistencies in texture statistics, frequency components, or generative artifacts.

4.2. Open Run (ResNet-50)

Training with ResNet-50 on the more diverse dataset improves generalization. Validation is more stable, and qualitative inspection of failures indicates fewer category-specific errors. The deeper model likely benefits from both increased capacity and exposure to varied content during training.

4.3. Generalization Strategies Explored

- **Cross-validation.** Using 5-fold CV provides more robust validation signals and helps detect overfitting that a single split may hide.
- **Data augmentation.** Lightweight augmentations reduce sensitivity to lighting, color, and local textures, modestly improving OOD behavior.
- **Maximizing labeled data.** Training on the full labeled set (when no validation is required) increases coverage and reduces variance.
- **Multiple checkpoints.** Running the same test set across several checkpoints/architectures offers ensemble-like insights and highlights variance across runs.

Tables 1 and 2 show results on validation and test set respectively. These results clearly show a performance gap between constrained run and open run. Constrained-run achieved low accuracy and AUC which led to poor generalization. In contrast, the Open-run performs a way better with much higher accuracy and ROC-AUC which interprets that increasing dataset variety substantially improves robustness and better generalization to unseen content.

Table 1
Results on Validation Set.

Run / Model	Accuracy	ROC-AUC	F1
Constrained (ResNet-18)	0.56	0.52	0.45
Open (ResNet-50)	0.91	0.97	0.91

Table 2
Results on Test Set.

Run / Model	Accuracy	ROC-AUC	F1
Constrained (ResNet-18)	0.46	0.44	0.41
Open (ResNet-50)	0.81	0.90	0.80

4.4. Unsuccessful Strategies

Before finalizing the ResNet-based framework, we systematically explored and evaluated several alternative strategies. These included deeper and variant CNN architectures (such as EfficientNet, DenseNet, and MobileNet), advanced data augmentations (Mixup, CutMix, Random Erasing), alternative loss functions (focal loss, label smoothing), and diverse optimization schemes (SGD, Lookahead, Ranger). We also experimented with domain adaptation and adversarial fine-tuning techniques. Across experiments, these methods often led to overfitting on the available training data, introduced optimization instability, or significantly increased computational cost without corresponding improvements in out-of-distribution validation or test performance. In some cases, overly aggressive augmentation degraded probability calibration, while pretraining on mismatched domains limited transfer effectiveness. Ultimately, a streamlined ResNet configuration—with moderate augmentation, cross-entropy loss, AdamW optimization, and careful data preprocessing—yielded the most stable and generalizable results.

5. Conclusion

- **Diversity over depth.** Model depth helps, but training distribution breadth is the primary driver of generalization.
- **Limited-domain training → OOD gap.** Models trained on narrowly scoped or domain-specific datasets exhibit limited generalization; when evaluated on broader or unseen image distributions, performance declines sharply.
- **Helpful but insufficient.** Augmentation and CV mitigate overfitting but cannot fully compensate for narrow data.

We compared constrained (ResNet-18) and open (ResNet-50) regimes for real vs. synthetic image detection in the MediaEval 2025 context. The constrained run generalized poorly due to category narrowness; the open run showed stronger transfer. Future work will explore transformer-based backbones, semi-supervised learning to leverage unlabeled data, and robustness techniques (e.g., frequency-domain features, adversarial training, or self-ensembling) to further improve OOD performance.

Declaration on Generative AI

The authors gratefully acknowledge the assistance of generative AI tools, including **ChatGPT** and **DeepSeek**, which were used to support idea formulation, code refinement, and language editing during the preparation of this paper. These tools were also employed in a limited capacity for generating synthetic image samples used in experimental data.

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, October 25–26, 2025. URL: https://multimediaeval.github.io/editions/2025/tasks/synthim/

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] A. Dosovitskiy, et al., An image is worth 16x16 words: Transformers for image recognition at scale, International Conference on Learning Representations (ICLR), 2021. URL: https://arxiv.org/abs/2010.11929.

[4] M. G. Ljungqvist, O. Nordander, M. Skans, A. Mildner, T. Liu, P. Nugues, Object Detector Differences when using Synthetic and Real Training Data, Proceedings of the 14th International Conference on Computer Vision Systems (ICVS), CEUR Workshop Proceedings, 2023. URL: http://ceur-ws.org/.

[5] U. Ojha, et al., Towards Universal Fake Image Detectors that Generalize Across Generative Models, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[6] Y. Zhang, et al., Unmasking AI-created visual content: a review of generated images and deepfake detection technologies, Journal of King Saud University - Computer and Information Sciences, vol. 37, p. 148, 2025. DOI: https://doi.org/10.1016/j.jksuci.2024.01.148.

[7] O. Li, et al., Improving Synthetic Image Detection Towards Generalization: An Image Transformation Perspective, arXiv preprint arXiv:2408.06741, 2024. URL: https://arxiv.org/abs/2408.06741.