Mattaru Net: Lightweight synthetic images detection using contrastive distilled learning

Ngoc-Tuan Nguyen[0], Phuc-Nguyen Lam-Gia[0], Nhat-Khai Hoang[0], Duc-Nghi Hoang[0] and Hai-Dang Nguyen[1]

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

Abstract

Modern synthetic image detectors often rely on data-dependent handcrafted features; even though some of the methods involving it yield high results, it requires case-specific engineering, hindering their practicality. Additionally, methods using these filters often employ heavy architectures with complicated aggregation mechanisms. MattaruNet is a model created specifically for tackling these problems. With that in mind, we introduce embed pinpointing, a distill-contrastive process that leverages an existing embedding space to anchor a new one, thereby reducing the search space for learning and facilitating faster convergence. Despite its minimal architecture, MattaruNet achieves 75.7% accuracy on the evaluation dataset, demonstrating competitive performance with significantly reduced complexity.

1. Introduction

Recent studies have provided useful insights and revealed gaps in detecting images generated by modern AI models [1, 2]. Historically, research in this sub-field focused on feature engineering—handcrafting filters to extract low-level artifacts—but this paradigm faces significant limitations. We contend that the scalability of methods relying on these handcrafted features is restricted: they are tedious to design, highly dataset-dependent, and often fail to generalize across different types of synthetic data [3]. Furthermore, deep learning models that aggregate multiple forensic cues often incorporate complex, long fusion mechanisms that significantly reduce speed, making them inefficient for large-scale classification.

Meanwhile, advances have been made in the field of computer vision. Such models, trained extensively on real-world images, are expected to encode rich embeddings of real-world structure. These embeddings as a reference would provide useful directions for learning synthetic features. Building on this idea, we design a model employing contrastive distilled learning [4] to learn presentations useful for detection, guided by the pretrained model's embedding. This encourages the student to replicate real-world representations while distinguishing synthetic content and improving generalization across both real and synthetic data.

In the 2025 MediaEval challenge, we set our sights on the Synthetic Images Task [5]: Advancing detection of generative AI used in real-world online images. The task is centered on developing AI models capable of detecting synthetic images. Our participation focuses on building a lightweight model to efficiently distinguish real images from their synthetic counterparts, achieving results comparable to networks of the same complexity. In this work, we

MediaEval'25: Multimedia Evaluation Workshop

k4deapricot.std@gmail.com (N. Nguyen); rawsienguyen@gmail.com (P. Lam-Gia); hoangnhatkhai2711@gmail.com (N. Hoang); nghihoang@proton.me (D. Hoang); nhdang@selab.hcmus.edu.vn (H. Nguyen)

CEUR Workshop Proceedings (CEUR-WS.org)

demonstrate that distillation learning effectively reduces model complexity while maintaining strong performance. We further show that contrastive learning allows the model to efficiently pinpoint relevant features, leading to faster convergence. Combining these ideas, we develop MattaruNet, a minimal yet effective detector that achieves 75.7% accuracy on the evaluation dataset. Our findings highlight the potential of lightweight, distill-contrastive approaches for advancing practical and scalable synthetic images detection.

2. Related Work

Early efforts in image forensics focused on handcrafted feature design and filters that expose camera or processing traces. These methods offered initial robustness by leveraging subtle artifacts. However, deep learning models developed to combine these forensic cues often incorporated complex fusion mechanisms, leading to architectures that are slow and poorly generalized against purely synthetic content [6, 3]. Over the last few years, the field has shifted toward using vanilla deep architectures - in particular ViT-style backbones [7]- because their self-attention layers naturally model long-range context and subtle global inconsistencies that handcrafted filters can miss. Works such as TransForensics [8], ProFact [9], and more recent transformer-first detectors (e.g., FatFormer [10]) demonstrate that transformer-based designs (including near 'pure' transformer backbones) can match or exceed hybrid CNN+ forensic pipelines.

In the broader field of computer vision, distillation, in which a teacher model guides the learning of a smaller model, has become a viable option to build lightweight and efficient networks, as demonstrated in TinyViT [11]. Additionally, contrastive learning, which encourages a model to cluster embeddings of the same class while dispersing opposites, helps the model learn consistent features, as used in SimCLR [4].

3. Methodology

3.1. Student - Teacher architecture with classification head

Our training setup comprises 2 models in parallel: a pre-trained teacher backbone and a tiny student model consists of 2 convolutional and a dense layers, as shown in Figure 1. The backbone training on a huge dataset of real images should have learned rich real semantics that could act as a reference for the real image subspace and form a transferable basis for learning the fake image subspace. As for the backbone, we tested our model in combination with multiple children from the model family EfficientNet [12] pre-trained on the ImageNet [13] dataset: B0, B1 and B2 to be exact. The student is a compact CNN model, designed to be extensively computationally efficient.

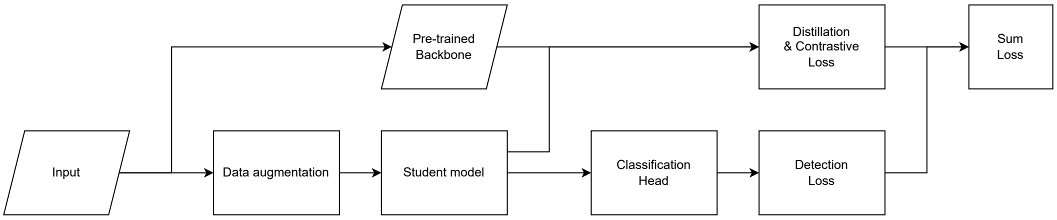

Figure 1: A simplified graph representing Mattaru model architecture.

3.2. Distillation, contrastive and classification loss

To stimulate interaction between these 2 networks, we employed 2 types of losses: the distillation loss allows the student model to learn embedding of reals image samples, reducing false positive classification; contrastive loss, on the other hand, encourages the student to learn embeddings diverting from the real base, enclosing a smaller solution space, and allows the model to effectively learn while being minimal. These 2 losses, however, only guide the model; it doesn't help the model to learn meaningful embeddings necessary for classification. Which is the reason a classification head is attached to the student with binary cross-entropy loss.

The overall loss is a weighted combination of the contrastive distillation loss and the classification loss. The alpha hyperparameter controls the contribution and is kept constant for MattaruB0, MattaruB1 and MattaruB2. It is multiplicatively reduced by a scheduler after each epoch for MattaruB2V2, however:

$$L = \alpha \cdot L_{contrast} + (1 - \alpha) \cdot L_{classification} \quad \{0.0 \leq \alpha \leq 1.0\}$$

3.3. Data augmentation

To prevent the model from learning trivial noise or local patterns inconsistent across different generative models or editing methods, we apply dynamic, runtime data augmentations during training. For all of the three runs, we use horizontal random flip, random rotation, random zoom, random translation, random brightness, random contrast, random saturation, and Gaussian noise, all scaled by a single hyperparameter for simplicity.

4. Experimental result

The evaluation results of the students trained by EfficientNetB0 [12], EfficientNetB1 [12], and EfficientNetB2 [12] are presented in Table 1.

Table 1
Results yielded by models trained with various teachers on the training set evaluated on the evaluation set, which both are provided by the organizers.

Model	Accuracy	Real Precision	Real Recall	Fake Precision	Fake Recall	F1 Score
CNN-2	51.30 %	49.90 %	47.90 %	52.51 %	54.50 %	53.49 %
MattaruNetB0	67.21%	61.90 %	**76.18** %	71.17 %	55.63 %	62.45 %
MattaruNetB1	**75.70**%	**77.23** %	72.90 %	75.66 %	**79.66** %	**77.61** %
MattaruNetB2	74.36%	72.88 %	75.37 %	**75.92** %	73.46 %	74.67 %

The testing results of the students trained by EfficientNetB0 [12], EfficientNetB1 [12], and EfficientNetB2 [12] are presented in Table 2.

Table 2
Results yielded by models trained with various teachers on the training set evaluated on the testing set, which both are provided by the organizers.

Model	Accuracy	Real Precision	Real Recall	Fake Precision	Fake Recall	F1 Score
MattaruNetB0	**53.07**%	**52.51** %	64.16 %	**53.95**%	**41.98**%	**47.22**%
MattaruNetB1	49.73%	49.80 %	**66.02** %	49.60%	33.44%	39.95%
MattaruNetB2	48.70%	49.02 %	65.28 %	48.06%	32.12%	38.50%

The results of Table 1 suggest that the models are able to fit the dataset and converge efficiently comparing to its peer, a 2-layer convolutional neural network. However, the result of Table 2, specifically the accuracy and precision, it is concluded that the model is struggling to generalize outside of the dataset it is trained on. Moreover, the high recall for real samples suggests that it is biased toward classifying a sample as real. Lastly, we suspect that B1 and B2 performed strongly on the evaluation set collapse on the test set due to resolution mismatch between the student and teacher as EfficientNet scaled; while B0 struck the relatively optimal balance and generalized best on the test set.

5. Discussion and Outlook

In general, it can be concluded that its learning is heavily restricted by its minimal capacity. As a result, the model collapses into learning noises specific to the dataset. Secondly, there is bias toward classifying a sample as real, we hypothesize that this bias comes from the unbalanced source of learning directions: The model is provided with meaningful embeddings with real samples to learn, while for the fake samples, it only uses the classification head and contrast to learn. Moreover, the formula for contrastive and distillation loss uses both cosine similarity and euclidean distance. While it is normalized, the operations are vastly different; As a result, learning is very unstable, and over time, the models fail to generalize well.

To address these limitations, we have to obviously scale the models up for further directions; Alternatively, another idea would be an architecture that employs two independent DINO [14] networks to separately learn embeddings for real and fake samples and use a more continuous loss function to fuse their feature together. However, this might inherit the collapse problem typical to adversarial networks as the two networks do not learn at the same rate.

6. Declaration on Generative AI

While writing this paper, the authors used ChatGPT as a tool to refine the wording and paraphrase some sentences to improve clarity; sentences which are unclear were discarded and rewritten by a human author. All information were kept inside a document as a checklist to ensure that no content was omitted. The authors also used Grammarly solely for spellings suggestions; Grammarly was not used for rephrasing. The author takes full responsibility of the paper's content.

References

[1] D. Karageorgiou, Q. Bammey, V. Porcellini, B. Goupil, D. Teyssou, S. Papadopoulos, Evolution of detection performance throughout the online lifespan of synthetic images, in: Trust What You learN (TWYN) Workshop ECCV 2024, 2024.

[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, ACM, 2024, pp. 55–64.

[3] F. Guillaro, G. Zingarini, B. Usman, A. Sud, D. Cozzolino, L. Verdoliva, A bias-free training paradigm for more general ai-generated image detection, in: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025.

[4] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, 2020.

[5] 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.

[6] D. Karageorgiou, S. Papadopoulos, I. Kompatsiaris, E. Gavves, Any-resolution ai-generated image detection by spectral learning, arXiv preprint arXiv:2411.19417 (2024).

[7] A. Dosovitskiy, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[8] J. Hao, Z. Zhang, S. Yang, D. Xie, S. Pu, Transforensics: Image forgery localization with dense self-attention, arXiv preprint arXiv:2108.03871 (2021).

[9] H. Zhu, G. Cao, X. Huang, Progressive feedback-enhanced transformer for image forgery localization (profact), arXiv preprint arXiv:2311.08910 (2023).

[10] H. Liu, Z. Tan, C. Tan, Y. Wei, Y. Zhao, J. Wang, Forgery-aware adaptive transformer for generalizable synthetic image detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[11] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, L. Yuan, Tinyvit: Fast pretraining distillation for small vision transformers, in: European Conference on Computer Vision (ECCV), Springer, 2022, pp. 68–85.

[12] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: ICML, 2019.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009) 248–255.

[14] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021).