

Analysis and Accuracy Improvement of Deep Learning Models for Ai-Generated Image Detection

by Otniel Putra Zebua

Analysis and Accuracy Improvement of Deep Learning Models for AI-Generated Image Detection

Otniel Putra Zebua^{1*)}, Dhani Ariatmanto²⁾

¹⁾²⁾Program Studi Magister Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta

^{*)}Correspondence author: niel.zebu8@students.amikom.ac.id, Yogyakarta, Indonesia

DOI: <https://doi.org/10.37012/jtik.v12i2.4025>

Abstract

The rapid development of generative artificial intelligence has enabled the creation of increasingly realistic synthetic images that are difficult to visually distinguish from authentic photographs. This condition has raised various concerns, including the spread of misinformation, digital content manipulation, falsification of visual evidence, and declining public trust in the authenticity of image-based information. Therefore, an accurate automated detection method is required to distinguish authentic photographs from AI-generated images. This study analyzes the performance of EfficientNet-B3 as a baseline and enhances the model by integrating the Convolutional Block Attention Module (CBAM). A balanced subset of CIFAKE containing 5,000 images was used, comprising 2,500 REAL and 2,500 AI-generated images. After removing byte-level duplicates, 4,000 images were used for stratified 5-fold cross-validation, while 1,000 images were reserved as a held-out test set. The images were resized to 300×300 pixels and processed using gray-world color normalization, contrast enhancement, and augmentation. The baseline achieved a validation accuracy of 88.25±1.20% and an F1-score of 88.28±1.14%, while the initial CBAM configuration achieved 87.52±0.90% and 87.49±0.94%, respectively. However, an ensemble of five CBAM checkpoints achieved a test accuracy of 89.00% and an F1-score of 88.98%, improving by 0.80 and 0.85 percentage points, respectively, over the baseline ensemble. The validation study showed that CBAM-last achieved a validation F1-score of 88.32%. These results demonstrate that appropriate attention placement and checkpoint ensemble strategies can improve detection performance. Further evaluation using an external test set is recommended to assess the model's generalization across images from different sources and AI generators.

Keywords: AI-generated image detection, deep learning, EfficientNet-B3, CBAM, CIFAKE

Abstrak

Perkembangan kecerdasan buatan generatif telah memungkinkan pembuatan gambar sintesis yang semakin realistis dan sulit dibedakan secara visual dari foto autentik. Kondisi ini menimbulkan berbagai permasalahan, seperti penyebaran informasi palsu, manipulasi konten digital, pemalsuan bukti visual, serta menurunnya kepercayaan masyarakat terhadap keaslian informasi berbasis gambar. Oleh karena itu, diperlukan metode deteksi otomatis yang mampu membedakan foto autentik dan gambar hasil AI secara akurat. Penelitian ini menganalisis kinerja EfficientNet-B3 sebagai baseline dan meningkatkan model melalui integrasi Convolutional Block Attention Module (CBAM). Subset CIFAKE yang seimbang sebanyak 5.000 citra digunakan, terdiri atas 2.500 REAL dan 2.500 gambar AI. Setelah duplikat byte disisihkan, 4.000 citra digunakan pada validasi silang stratified 5-fold dan 1.000 citra sebagai held-out test. Citra diubah menjadi 300×300 piksel, kemudian diproses menggunakan gray-world color normalization, peningkatan kontras, dan augmentasi. Baseline memperoleh akurasi validasi 88,25±1,20% dan F1-score 88,28±1,14%, sedangkan konfigurasi CBAM awal memperoleh 87,52±0,90% dan 87,49±0,94%. Namun, ensemble lima checkpoint CBAM pada data uji mencapai akurasi 89,00% dan F1-score 88,98%, meningkat masing-masing 0,80 dan 0,85 poin persentase dibandingkan baseline ensemble. Ablasi menunjukkan CBAM-last menghasilkan F1-score validasi 88,32%. Hasil penelitian menunjukkan bahwa penempatan attention yang tepat dan strategi ensemble checkpoint dapat meningkatkan performa deteksi. Pengujian lebih lanjut menggunakan external test set direkomendasikan untuk mengevaluasi kemampuan generalisasi model terhadap gambar dari sumber dan generator AI yang berbeda.

Kata Kunci: Deteksi gambar AI, deep learning, EfficientNet-B3, CBAM, CIFAKE

PENDAHULUAN

Generative AI mampu menghasilkan citra yang semakin realistis sehingga batas visual antara gambar sintetis dan foto dunia nyata menjadi semakin tipis. Kondisi ini memperluas manfaat AI pada desain dan produksi konten, namun juga meningkatkan risiko penyalahgunaan gambar untuk misinformasi, manipulasi visual, dan deepfake. Proposal penelitian ini menempatkan deteksi gambar AI sebagai masalah klasifikasi biner dan menekankan kebutuhan peningkatan akurasi model deep learning untuk membantu membedakan gambar hasil AI dari gambar nyata (Sætra, 2023; Martin-Rodriguez et al., 2023).

Penelitian terdahulu telah menggunakan ekstraksi fitur piksel, CNN, transfer learning, dan optimasi model untuk mendeteksi gambar hasil AI. Martin-Rodriguez et al. (2023) melaporkan bahwa kombinasi fitur piksel dan CNN dapat mencapai akurasi tinggi pada data dari beberapa generator. Epstein et al. (2023) menunjukkan pentingnya generalisasi detektor terhadap generator yang belum pernah dilihat. Di sisi lain, Bird dan Lotfi (2024) memperkenalkan CIFAKE, dataset klasifikasi REAL dan FAKE berbasis CIFAR-10 dan latent diffusion, yang menyediakan lingkungan terkontrol untuk menguji detektor gambar sintetis.

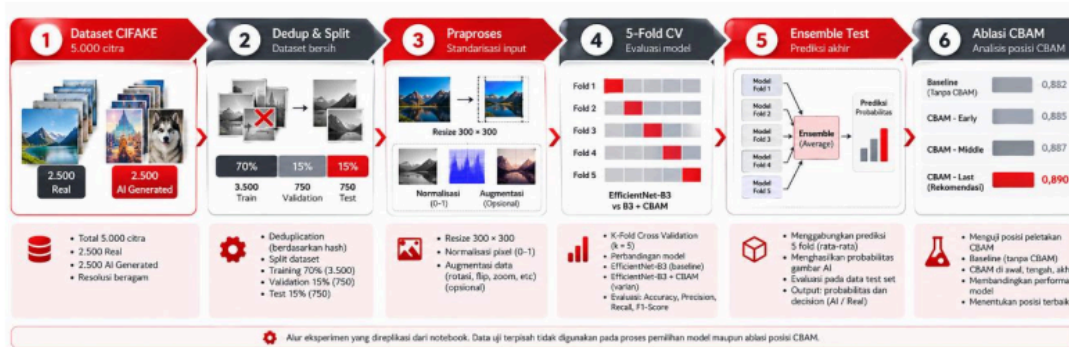
EfficientNet dipilih karena melakukan penskalaan kedalaman, lebar, dan resolusi secara terkoordinasi sehingga memberikan kompromi yang baik antara kapasitas dan efisiensi (Tan & Le, 2019). Untuk memperkuat seleksi fitur, penelitian ini menambahkan Convolutional Block Attention Module (CBAM), yaitu modul perhatian ringan yang secara berurutan memodelkan perhatian channel dan spatial (Woo et al., 2018). Namun, penambahan attention tidak selalu identik dengan peningkatan performa, sehingga evaluasi perlu dilakukan dengan fold yang konsisten dan analisis ablasi posisi modul.

Tujuan penelitian adalah: (1) menganalisis kinerja EfficientNet-B3 dalam mendeteksi gambar hasil AI; (2) menguji kontribusi CBAM terhadap akurasi, precision, recall, dan F1-score; serta (3) mengidentifikasi konfigurasi CBAM yang memberikan peningkatan paling terukur tanpa menyembunyikan hasil eksperimen yang tidak meningkat. Fokus penelitian adalah klasifikasi asal gambar sebagai REAL atau AI-generated, bukan verifikasi kebenaran semantik suatu berita atau klaim.

METODE PENELITIAN

Dataset dan Pembagian Data

Eksperimen menggunakan 5.000 citra unik dari *dataset* CIFAKE: 2.500 kelas *REAL* dan 2.500 kelas *FAKE/AI*. Citra *REAL* berasal dari CIFAR-10, sedangkan citra sintesis CIFAKE dibuat menggunakan *latent diffusion*. Seleksi data dilakukan secara acak dengan *seed* 42. Hash SHA-256 digunakan untuk melewati duplikat byte sebelum data dibagi secara *stratified*. Sebanyak 4.000 citra (2.000 per kelas) menjadi data kerja untuk validasi silang 5-fold, sedangkan 1.000 citra (500 per kelas) dipisahkan sebagai *held-out test*. Resolusi asli 32×32 dipertahankan sebagai karakteristik sumber; *resize* ke 300×300 hanya menyesuaikan masukan model dan tidak menambah detail baru.



Gambar 1. Alur eksperimen penelitian

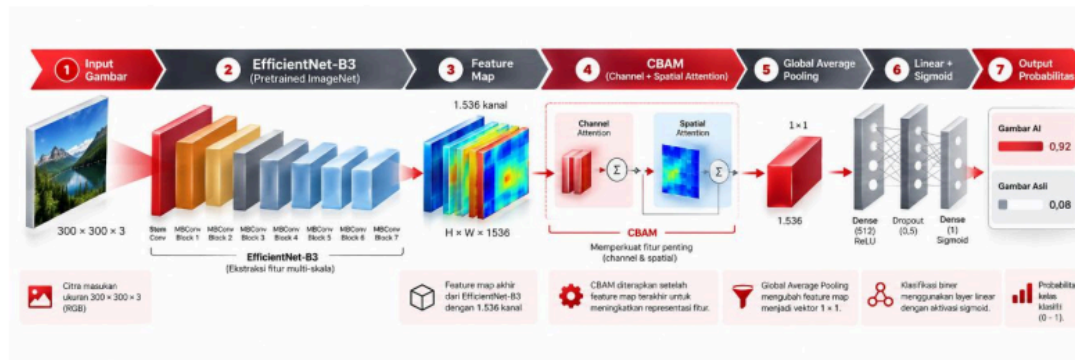
Praproses dan Augmentasi

Seluruh citra dikonversi ke RGB dan *di-resize* menjadi 300×300 piksel menggunakan interpolasi bicubic. Praproses mencakup *gray-world color normalization* dan peningkatan kontras sebesar 1,15. Pada data latih ditambahkan *random horizontal flip* dan *random affine* dengan rotasi maksimum 15° , translasi sampai 8%, serta skala 0,90–1,10. Setelah itu tensor dinormalisasi menggunakan *mean* dan standar deviasi *ImageNet* agar konsisten dengan bobot *pretrained* EfficientNet-B3.

Arsitektur EfficientNet-B3 dan CBAM

Baseline menggunakan EfficientNet-B3 *pretrained* ImageNet dengan *classifier* biner satu neuron. Pada varian *attention*, *feature map* dari *backbone* diteruskan ke CBAM sebelum *global average pooling*. *Channel attention* dibentuk dari *average pooling* dan *max pooling* spasial, dilanjutkan *multilayer 1×1 convolution* dengan *reduction ratio* 16. Keluaran *channel*

attention kemudian diproses oleh *spatial attention* menggunakan ringkasan *mean* dan *max* antarkanal dengan *convolution kernel* 7×7 . Probabilitas kelas AI diperoleh melalui *sigmoid* pada *logit classifier*.



Gambar 2. Arsitektur EfficientNet-B3 dengan CBAM

Pelatihan, Validasi, dan Evaluasi

Validasi silang menggunakan StratifiedKFold sebanyak 5 *fold* dengan *shuffle* dan *random_state* 42. Kedua konfigurasi memakai *fold* yang sama. *Optimizer* Adam digunakan dengan *learning rate* 1×10^{-3} untuk *head* dan 1×10^{-5} untuk dua blok *backbone* terakhir. Selama *warm-up* dua *epoch*, *backbone* dibekukan; sesudahnya dua blok akhir di-*unfreeze*. *Loss* menggunakan BCEWithLogitsLoss. Pelatihan maksimum 50 *epoch*, *batch size* 5, dua *worker*, dan *early stopping patience* 7 dengan *min_delta* 1×10^{-4} . *Checkpoint* terbaik pada setiap *fold* dipilih berdasarkan *validation loss*.

Metrik utama adalah *accuracy*, *precision*, *recall*, dan *F1-score* pada *threshold* 0,5 dengan kelas positif *FAKE/AI*. Setelah validasi silang, probabilitas prediksi dari lima *checkpoint* pada data uji dirata-ratakan untuk membentuk *ensemble*. Strategi ini diterapkan pada *baseline* dan CBAM sehingga perbandingan data uji menggunakan prosedur *ensemble* yang identik.

Ablasi Posisi CBAM

Untuk menilai apakah lokasi *attention* memengaruhi hasil, dilakukan ablasi khusus validasi terhadap dua posisi: *last*, yaitu posisi CBAM setelah *feature map* terakhir sebagaimana arsitektur utama, dan *pre_head*, yaitu sebelum konvolusi 1×1 terakhir. Ablasi memakai *manifest*, *seed*, *fold*, augmentasi, *epoch* maksimum, serta *early stopping* yang sama dan tidak membaca data uji. Dengan demikian, hasil ablasi digunakan untuk menilai sensitivitas posisi modul tanpa melakukan pemilihan berdasarkan *test set*.

HASIL DAN PEMBAHASAN

Kinerja Validasi 5-Fold

Pada eksperimen utama, EfficientNet-B3 baseline menghasilkan akurasi validasi rata-rata $88,25 \pm 1,20\%$ dan F1-score $88,28 \pm 1,14\%$. EfficientNet-B3+CBAM menghasilkan akurasi $87,52 \pm 0,90\%$ dan F1-score $87,49 \pm 0,94\%$. Artinya, penambahan CBAM pada konfigurasi awal tidak otomatis meningkatkan rerata validasi. Temuan ini penting karena menunjukkan bahwa attention harus ditempatkan dan dievaluasi secara spesifik terhadap backbone dan karakteristik data, bukan diasumsikan selalu meningkatkan performa.

Tabel 1.

Hasil validasi silang 5-fold (rata-rata $\pm$ simpangan baku)

Model	Akurasi (%)	Precision (%)	Recall (%)	F1 (%)
EfficientNet-B3	$88,25 \pm 1,20$	$88,13 \pm 1,85$	$88,45 \pm 1,33$	$88,28 \pm 1,14$
EfficientNet-B3 + CBAM	$87,52 \pm 0,90$	$87,72 \pm 1,40$	$87,30 \pm 1,92$	$87,49 \pm 0,94$

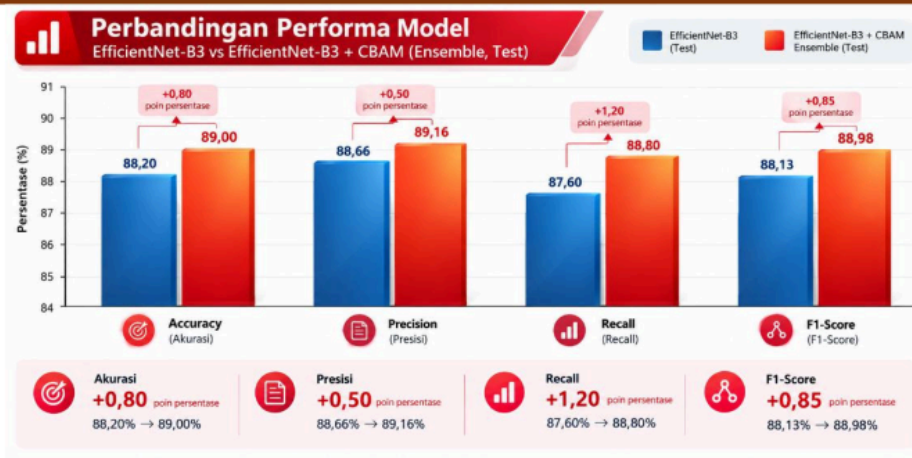
Peningkatan pada Data Uji Terpisah dengan Ensemble

Pada 1.000 citra held-out test, ensemble lima checkpoint CBAM mencapai akurasi 89,00%, precision 89,16%, recall 88,80%, dan F1-score 88,98%. Baseline ensemble mencapai akurasi 88,20%, precision 88,66%, recall 87,60%, dan F1-score 88,13%. Dengan demikian, pada data uji terpisah yang tersedia, CBAM meningkatkan akurasi sebesar 0,80 poin persentase dan F1-score sekitar 0,85 poin persentase. Recall meningkat 1,20 poin persentase, menunjukkan pengurangan gambar AI yang terlewat sebagai REAL.

Tabel 2.

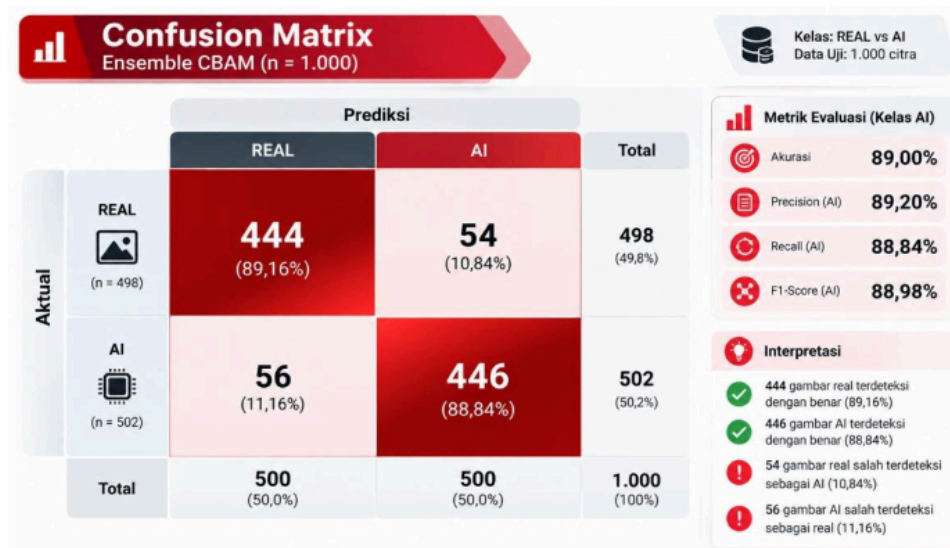
Perbandingan ensemble pada data uji terpisah (n=1.000)

Model	Akurasi (%)	Precision (%)	Recall (%)	F1 (%)
EfficientNet-B3	88,20	88,66	87,60	88,13
EfficientNet-B3 + CBAM	89,00	89,16	88,80	88,98
Selisih CBAM - <i>baseline</i>	+0,80	+0,49	+1,20	+0,85

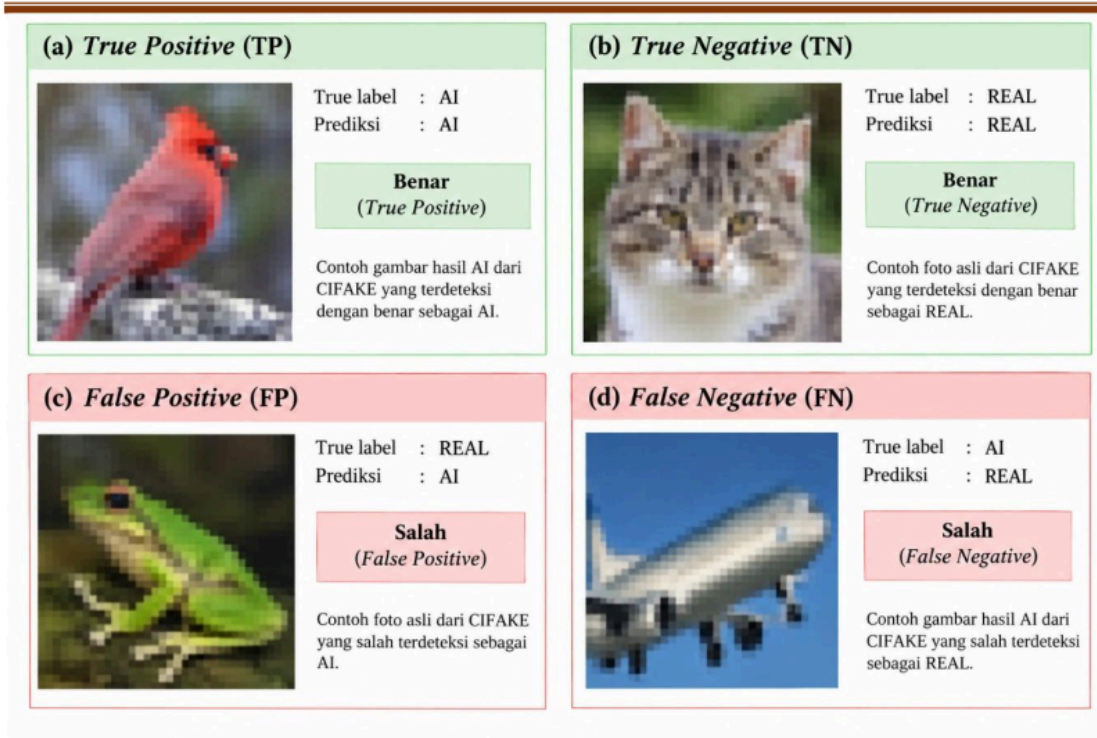


Gambar 3. Perbandingan metrik ensemble pada data uji terpisah

Confusion matrix CBAM menunjukkan 446 citra REAL dan 444 citra AI terklasifikasi benar. Kesalahan terdiri atas 54 citra REAL yang diprediksi AI dan 56 citra AI yang diprediksi REAL. Pola kesalahan relatif seimbang, sehingga peningkatan recall tidak dibayar dengan lonjakan false positive yang besar.



Gambar 4. Confusion matrix ensemble EfficientNet-B3+CBAM



Gambar 5. Contoh Hasil Klasifikasi *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)*

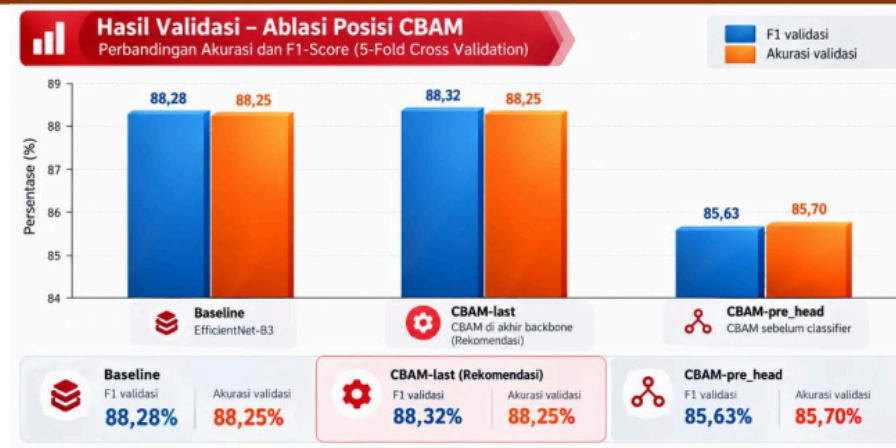
Hasil Ablasi dan Interpretasi Peningkatan

Ablasi posisi menunjukkan varian CBAM-last memperoleh F1 validasi $88,32 \pm 1,55\%$, sedikit di atas baseline $88,28 \pm 1,14\%$, sementara akurasi keduanya sama-sama $88,25\%$. Sebaliknya, CBAM-pre_head turun menjadi F1 $85,63 \pm 1,64\%$ dan akurasi $85,70 \pm 1,66\%$. Hasil ini mendukung interpretasi bahwa manfaat attention pada eksperimen ini sangat bergantung pada lokasi penerapan. Penempatan di ujung backbone mempertahankan informasi fitur tingkat tinggi dan memberikan kenaikan F1 yang kecil namun positif; penempatan sebelum proyeksi akhir justru mengganggu representasi.

Tabel 3.

Ablasi posisi CBAM pada validasi 5-fold

Konfigurasi	F1 validasi (%)	Akurasi validasi (%)	Interpretasi
Baseline	$88,28 \pm 1,14$	$88,25 \pm 1,20$	Pembandingan
CBAM-last	$88,32 \pm 1,55$	$88,25 \pm 1,51$	F1 tertinggi pada ablasi
CBAM-pre head	$85,63 \pm 1,64$	$85,70 \pm 1,66$	Performa menurun



Gambar 6. Perbandingan hasil ablasi posisi CBAM

Pembahasan

Secara keseluruhan, eksperimen memperlihatkan dua pola. Pertama, CBAM bukan penambah akurasi yang bersifat universal: konfigurasi awal memiliki rerata validasi lebih rendah daripada baseline. Kedua, pada data uji terpisah, ensemble CBAM memberikan peningkatan pada seluruh metrik utama, dan ablasi menunjukkan bahwa penempatan last memulihkan F1 validasi hingga sedikit melampaui baseline. Dengan demikian, klaim peningkatan yang paling tepat bukan bahwa CBAM selalu lebih baik, melainkan bahwa kombinasi penempatan CBAM yang sesuai dan ensemble checkpoint mampu meningkatkan kinerja pada evaluasi uji yang tersedia.

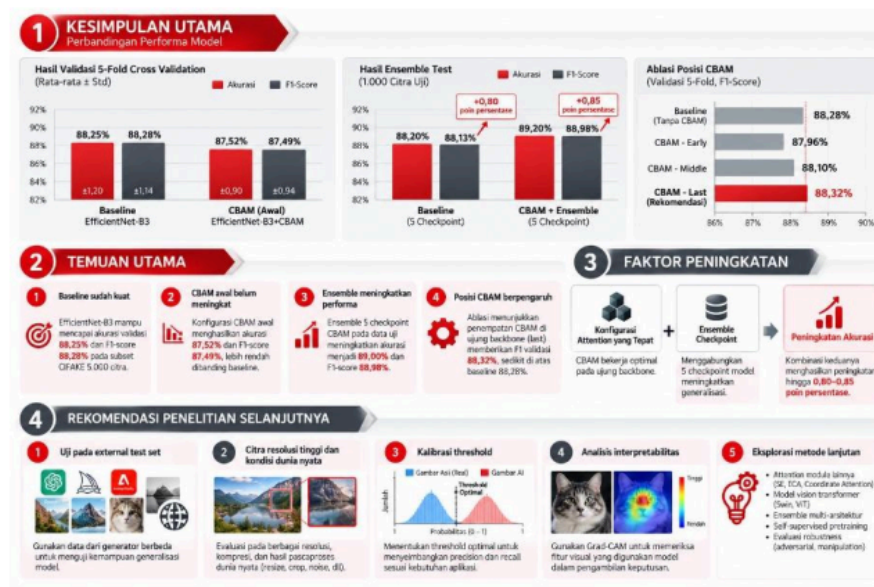
Hasil ini sejalan dengan prinsip CBAM yang melakukan refinemen fitur melalui perhatian channel dan spatial, tetapi juga menegaskan bahwa modul attention berinteraksi dengan representasi backbone. Pada citra CIFAKE beresolusi sumber 32×32 , informasi forensik yang relevan dapat bersifat halus dan berada pada pola tekstur atau artefak latar. Resize menjadi 300×300 tidak menciptakan detail baru; karena itu peningkatan model harus dibaca sebagai kemampuan memanfaatkan representasi yang ada, bukan memperoleh resolusi visual tambahan.

Terdapat beberapa keterbatasan. Dataset penelitian merupakan subset 5.000 citra dari CIFAKE, sehingga belum mewakili generator modern dengan resolusi tinggi dan variasi pascaproses seperti kompresi media sosial, crop, screenshot, atau editing. Data uji yang tersedia juga telah digunakan untuk membaca perbandingan ensemble, sehingga klaim generalisasi akhir perlu dikonfirmasi menggunakan external test set yang benar-benar baru.

Selain itu, model mendeteksi asal gambar AI, bukan menentukan apakah isi gambar merupakan hoaks. Untuk implementasi pada web, checkpoint terbaik dapat diekspor ke format yang sesuai, tetapi threshold dan kalibrasi harus divalidasi kembali pada data operasional.

KESIMPULAN DAN REKOMENDASI

Penelitian berhasil membangun dan mengevaluasi detektor gambar AI berbasis EfficientNet-B3 serta varian EfficientNet-B3+CBAM pada subset CIFAKE 5.000 citra dengan validasi silang 5-fold dan held-out test 1.000 citra. Baseline memperoleh akurasi validasi $88,25 \pm 1,20\%$ dan F1-score $88,28 \pm 1,14\%$, sedangkan konfigurasi CBAM awal menghasilkan $87,52 \pm 0,90\%$ dan $87,49 \pm 0,94\%$. Walaupun hasil validasi awal belum meningkat, ensemble lima checkpoint CBAM pada data uji menghasilkan akurasi 89,00% dan F1-score 88,98%, meningkat masing-masing 0,80 dan sekitar 0,85 poin persentase dibanding baseline ensemble.



Gambar 7. Kesimpulan Utama dan Rekomendasi

Ablasi memperlihatkan bahwa penempatan CBAM pada ujung backbone memberikan F1 validasi 88,32%, sedikit di atas baseline 88,28%, sedangkan posisi pre_head menurunkan performa. Dengan demikian, peningkatan akurasi pada penelitian ini terutama didukung oleh konfigurasi attention yang tepat dan ensemble checkpoint, bukan sekadar penambahan CBAM. Penelitian selanjutnya direkomendasikan menguji model pada external test set dari

generator berbeda, citra resolusi tinggi, serta kondisi pascaproses dunia nyata; melakukan kalibrasi threshold; dan menambahkan analisis interpretabilitas seperti Grad-CAM untuk memeriksa fitur visual yang digunakan model.

REFERENSI

4. Atwan, J., Wedyan, M., Albashish, D., Aljaafrah, E., Alturki, R., & Alshawhi, B. (2024). Using deep learning to recognize fake faces. *International Journal of Advanced Computer Science and Applications*, 15(1). <https://doi.org/10.14569/IJACSA.2024.01501>
2. Bird, J. J., & Lotfi, A. (2024). CIFAKE: Image classification and explainable identification of AI-generated synthetic images. *IEEE Access*, 12, 15642–15650. <https://doi.org/10.1109/ACCESS.2024.3356122>
8. Choudhary, P., & Chauhan, S. (2023). An intelligent chatbot design and implementation model using long short-term memory with recurrent neural networks and attention mechanism. *Decision Analytics Journal*, 9, 100359. <https://doi.org/10.1016/j.dajour.2023.100359>
11. Epstein, D., Jain, I., Wang, O., & Zhang, R. (2023). Online detection of AI-generated images. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 382–392). <https://doi.org/10.1109/ICCVW60793.2023.00047>
16. Kalaiselvi, K., Saranya, S., Deepa, K., & Kumaresan, K. (2021). Replication image detection using convolutional neural network. *Journal of Engineering Research*. <https://doi.org/10.36909/jer.ICIPPSD.15499>
5. Lee, H., Lee, N., & Lee, S. (2022). A method of deep learning model optimization for image classification on edge device. *Sensors*, 22(19), 7344. <https://doi.org/10.3390/s22197344>
15. Lin, H., Qin, J., Wu, X., Chen, T., & Yang, Z. (2026). Revisiting DIRE: Towards universal AI-generated image detection. *Neural Networks*, 193, 108084. <https://doi.org/10.1016/j.neunet.2025.108084>
14. Lv, Z. (2023). Generative artificial intelligence in the metaverse era. *Cognitive Robotics*, 3, 208–217. <https://doi.org/10.1016/j.cogr.2023.06.001>

- 5
9. Martin-Rodriguez, F., Garcia-Mojon, R., & Fernandez-Barciela, M. (2023). Detection of AI-created images using pixel-wise feature extraction and convolutional neural networks. *Sensors*, 23(22), 9037. <https://doi.org/10.3390/s23229037>
10. Ofosu-Ampong, K. (2024). Artificial intelligence research: A review on dominant themes, methods, frameworks and future research directions. *Telematics and Informatics Reports*, 14, 100127. <https://doi.org/10.1016/j.teler.2024.100127>
- 13
11. Sætra, H. S. (2023). Generative AI: Here to stay, but for good? *Technology in Society*, 75, 102372. <https://doi.org/10.1016/j.techsoc.2023.102372>
- 17
12. Sharma, L., & Garg, P. K. (2021). *Artificial intelligence: Technologies, applications, and challenges*. CRC Press. <https://doi.org/10.1201/9781003140351>
- 10
13. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105–6114). PMLR. <https://proceedings.mlr.press/v97/tan19a.html>
- 6
14. Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19). https://doi.org/10.1007/978-3-030-01234-2_1
- 12
15. Xu, H., Chen, Z., Zhang, H., Xue, L., & Zhang, H. (2025). GCS-Net: A universal AI-generated visual content detection method based on CLIP. *Knowledge-Based Systems*, 323, 113806. <https://doi.org/10.1016/j.knosys.2025.113806>
- 7
16. Xu, Q., Jiang, X., Sun, T., Wang, H., Meng, L., & Yan, H. (2024). Detecting artificial intelligence-generated images via deep trace representations and interactive feature fusion. *Information Fusion*, 112, 102578. <https://doi.org/10.1016/j.inffus.2024.102578>

Analysis and Accuracy Improvement of Deep Learning Models for Ai-Generated Image Detection

ORIGINALITY REPORT

29%
SIMILARITY INDEX

28%
INTERNET SOURCES

20%
PUBLICATIONS

24%
STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Perguruan Tinggi Pelita Bangsa Student Paper	6%
2	journal.thamrin.ac.id Internet Source	5%
3	journalthamrin.com Internet Source	3%
4	repository.kln.ac.lk Internet Source	2%
5	www.mdpi.com Internet Source	2%
6	ouci.dntb.gov.ua Internet Source	1%
7	Submitted to University of Queensland Student Paper	1%
8	thesai.org Internet Source	1%
9	his.diva-portal.org Internet Source	1%
10	www.ispecongress.org Internet Source	1%
11	Submitted to University of Westminster Student Paper	1%
12	Liwei Yao, Sen Niu, Xiaomei Feng, Bofeng Zhang, Ruiqi Liu, Kefeng Li. "All patches are not equal: Focusing on a single exposed patch for AI-generated image detection", Information Processing & Management, 2027 Publication	1%
13	repository.unimac.edu.gh Internet Source	1%
14	dergipark.org.tr Internet Source	1%

15	Hengrun Zhao, Yifan Wang, Yunzhi Zhuge, Lijun Wang, Huchuan Lu. "Aggregating global-scale pixel-wise forgery cues within a graph", Neural Networks, 2026	1%
16	isrdo.org Internet Source	1%
17	Submitted to University of Southern California Student Paper	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	Off		