Deep Learning Framework for Automatic Tagging of Multimedia Content

Ikna Awaliyani*[1], Dita Septasari[2], Nur Aminudin[3], Ulfa Isni Kurnia[4]
[1,2]Pendidikan Teknologi Informasi, Universitas Aisyah Pringsewu, Lampung
[3]Rekayasa Perangkat Lunak, Universitas Aisyah Pringsewu, Lampung
e-mail: ***[1]iknaawaliyani@aisyahuniversity.ac.id**, [2]ditaseptasari@aisyahuniversity.ac.id,
[3]nuraminudin@aisyahuniversity.ac.id, [4]ulfaisnikurnia@aisyahuniversity.ac.id

Abstrak

Penandaan otomatis konten multimedia merupakan langkah penting dalam sistem pengelolaan aset digital, rekomendasi konten, dan pencarian berbasis visual. Namun, kompleksitas konteks gambar, keberagaman objek, serta karakteristik *multi-label* menjadi tantangan bagi metode konvensional. Penelitian ini mengusulkan sebuah *framework deep learning end to end berbasis Vision Transformer (ViT)* yang dikombinasikan dengan adaptive *multi-label classification head* serta opsi modul multimodal untuk memanfaatkan hubungan semantik antara gambar dan teks. Eksperimen dilakukan pada *dataset* berskala besar seperti MS-COCO, NUS-WIDE, dan *Open Images Dataset*, dengan strategi pelatihan mencakup augmentasi, *fine-tuning* bertahap, dan adaptive loss. Hasil pengujian menunjukkan bahwa model usulan mencapai peningkatan performa 4–6% dibandingkan *baseline* CNN dan 2–3% di atas *Vision Transformer* murni, dengan skor *mAP* 0,78 dan *F1-score* 0,82. Pendekatan multimodal memberikan peningkatan tambahan pada label semantik tinggi seperti activity, outdoor, dan event. Temuan ini menunjukkan bahwa integrasi *Vision Transformer* dan modul multimodal dapat meningkatkan akurasi dan konteks semantik dalam automatic tagging konten multimedia modern. Berbeda dengan pendekatan CNN konvensional dan *Vision Transformer* murni pada penelitian terdahulu, *framework* yang diusulkan mengintegrasikan adaptive *multi-label classification head* serta opsi *multimodal learning* secara *end to end*, yang terbukti mampu meningkatkan akurasi, stabilitas prediksi, dan pemahaman semantik pada skenario penandaan multimedia berskala besar.

Kata kunci*— Automatic Tagging, Deep Learning, Vision Transformer, Multimodal Learning, Multi-Label Classification.*

Abstract

Automatic multimedia content tagging is a critical component in digital asset management systems, content recommendation, and visual-based search. However, the complexity of visual context, object diversity, and the inherent multi-label characteristics pose significant challenges to conventional methods. This study proposes an end-to-end deep learning framework based on Vision Transformers (ViT), combined with an adaptive multi-label classification head and an optional multimodal module to exploit semantic relationships between images and textual information. Experiments were conducted on large-scale datasets, including MS-COCO, NUS-WIDE, and the Open Images Dataset, using training strategies such as data augmentation, progressive fine-tuning, and adaptive loss functions. The experimental results demonstrate that the proposed model achieves performance improvements of 4–6% over CNN-based baselines and 2–3% over a standard Vision Transformer, achieving a mean Average Precision (mAP) of 0.78 and an F1-score of 0.82. The multimodal approach provides additional gains for high-level semantic labels such as activity, outdoor, and event. These findings indicate that the integration of Vision Transformers and multimodal modules can significantly enhance both accuracy and semantic contextual understanding in modern automatic multimedia content tagging. Unlike conventional CNN-based approaches and standalone Vision Transformer models

reported in previous studies, the proposed framework integrates an adaptive multi-label classification head and optional end-to-end multimodal learning, which has been shown to improve accuracy, prediction stability, and semantic understanding in large-scale multimedia tagging scenarios.

Keywords— *Deep Learning, Multimedia Tagging, Vision Transformer, Multimodal Learning.*

1. PENDAHULUAN

Perkembangan teknologi digital dan masifnya penggunaan platform berbasis multimedia seperti *Instagram, YouTube,* dan *TikTok* menyebabkan volume data visual dan audiovisual meningkat secara eksponensial dalam satu dekade terakhir [1], [2], [3]. Pertumbuhan data ini menimbulkan kebutuhan mendesak terhadap mekanisme pengelolaan dan pengindeksan konten yang lebih efisien, khususnya dalam proses penandaan (*tagging*) yang menjadi fondasi bagi sistem rekomendasi, pencarian semantik, dan digital asset management. Tag yang akurat dan konsisten tidak hanya membantu pengelompokan konten, tetapi juga menentukan kualitas personalisasi dan relevansi dalam berbagai sistem layanan multimedia berskala besar [4], [5], [6]. Namun, sebagian besar sistem masih bergantung pada penandaan manual atau teknik konvensional yang tidak mampu memenuhi tantangan skalabilitas, kompleksitas konteks visual, serta keragaman objek dalam data modern.

Kemajuan deep learning, khususnya *Convolutional Neural Networks (CNN)*, telah memberikan terobosan besar pada tugas ekstraksi fitur visual berhirarki dan klasifikasi gambar berskala besar [7], [8]. Namun, keterbatasannya dalam menangkap dependensi global dan hubungan kontekstual antarpotongan gambar membuat CNN kurang optimal untuk tugas-tugas penandaan *multi-label* yang memerlukan pemahaman semantik lebih luas. Kemunculan *Vision Transformer* (ViT) menawarkan paradigma baru melalui mekanisme *self-attention* yang mampu mempelajari hubungan global secara lebih efektif [9], [10], [11], [12]. Selain itu, perkembangan model *multimodal* seperti CLIP menunjukkan bahwa integrasi representasi gambar dan teks dapat meningkatkan kualitas anotasi, terutama pada label abstrak seperti aktivitas atau konteks lingkungan [13], [14]. Meskipun berbagai pendekatan ini telah menunjukkan potensi besar, sebagian besar masih menghadapi tantangan pada ambiguitas label, ketergantungan pada *dataset* beranotasi manual, serta keterbatasan dalam menangani multimodalitas kompleks pada data berskala besar [15], [16].

Sejumlah penelitian juga menyoroti kebutuhan akan sistem yang mampu bekerja secara *end to end* dan efisien dari sisi komputasi. Model-model *vision-language* modern umumnya memiliki kompleksitas tinggi dan membutuhkan sumber daya besar, sehingga kurang ideal untuk implementasi *real time* pada platform industri [17], [18], [19]. Selain itu, banyak model belum sepenuhnya mampu beradaptasi dengan perubahan distribusi data (*distribution shift*), yang umum terjadi pada ekosistem multimedia dunia nyata [20]. Dengan demikian, masih terdapat gap penelitian terkait bagaimana merancang *framework deep learning* yang mampu memadukan pemahaman visual dan semantik secara komprehensif, tetap efisien dari sisi komputasi, serta adaptif terhadap berbagai variasi konten.

Berdasarkan kesenjangan tersebut, penelitian ini mengangkat tiga rumusan masalah utama: (1) bagaimana merancang sebuah *framework deep learning* yang mampu melakukan penandaan konten *multimedia* secara otomatis dan *end to end*; (2) bagaimana mengoptimalkan arsitektur agar tetap efisien dan adaptif terhadap data berskala besar dan multimodal; serta (3) bagaimana mengevaluasi performa model secara komprehensif pada konteks *multi-label*, distribusi data yang berubah, dan skenario dunia nyata yang beragam.

Penelitian ini bertujuan mengembangkan sebuah *framework* berbasis *Vision Transformer* dan modul multimodal untuk menghasilkan anotasi *multi-label* yang akurat dan relevan secara semantik. Pendekatan ini mengintegrasikan representasi visual dan teks dengan strategi pelatihan adaptif seperti *augmentasi, fine-tuning* bertahap, dan *adaptive loss* untuk mengatasi *class*

imbalance. Selain itu, penelitian ini berupaya meningkatkan efisiensi model melalui optimasi komputasi sehingga lebih layak untuk implementasi industri berskala besar.

2. METODE PENELITIAN

Metode penelitian ini dirancang untuk mengembangkan dan mengevaluasi sebuah *framework deep learning end to end* yang mampu melakukan penandaan otomatis (*automatic tagging*) terhadap konten multimedia, khususnya data citra dan citra bertautan teks. Penelitian ini bersifat eksperimental dan dilakukan melalui serangkaian tahapan mulai dari pengumpulan data, praproses, desain arsitektur model, pelatihan, hingga evaluasi kinerja.

2.1 Desain Penelitian

Penelitian ini menerapkan pendekatan eksperimen kuantitatif berbasis komputasi untuk pengembangan model deep learning dalam prediksi multi-label. Tahapan penelitian mencakup pemilihan dan persiapan dataset, praproses dan augmentasi data, perancangan arsitektur CNN/ViT atau multimodal, pelatihan dan optimasi model, serta evaluasi, analisis, dan validasi hasil.

Desain ini mengikuti praktik terbaik dalam penelitian *computer vision* dan *multimodal learning*.

2.2 Dataset Penelitian

2.2.1 Sumber Dataset

Penelitian ini menggunakan dataset berskala besar yang umum pada tugas penandaan citra *multi-label*, yaitu *MS-COCO, NUS-WIDE*, dan *Open Images Dataset (OID)*. Dataset tersebut dipilih karena menyediakan anotasi multi-label yang sesuai dengan kebutuhan pengembangan dan evaluasi sistem tagging otomatis.

2.2.2 Karakteristik Dataset

Dataset terdiri dari citra RGB dengan resolusi variatif dan anotasi *multi-label* (1–10 tag per gambar), dengan total lebih dari 100.000 sampel. *Dataset* juga menunjukkan ketidakseimbangan kelas yang signifikan, terutama pada label berfrekuensi rendah.

2.2.3 Pembagian Dataset

Dataset dibagi secara *stratified* menjadi *training (70%), validation (15%),* dan *test set (15%)* untuk menjaga proporsi label pada setiap *subset.*

2.3 Praproses dan Augmentasi Data

2.3.1 Praproses Citra

Pra-pemrosesan data meliputi *resize* citra ke 224×224 atau 384×384 piksel, normalisasi menggunakan mean dan standar deviasi ImageNet, serta *multi-hot encoding* untuk representasi label *multi-label*.

2.3.2 Augmentasi Citra

Untuk meningkatkan variasi data dan mencegah *overfitting*, diterapkan augmentasi berupa *random horizontal flip*, *random crop, color jitter, random rotation*, serta *Gaussian blur* (opsional).

2.3.3 Praproses Teks (Jika Multimodal)

Pada dataset dengan pasangan gambar–teks, *pra-pemrosesan* teks meliputi *lowercasing, tokenisasi*, penghapusan simbol *non-alfabet*, serta *padding* dan *masking*.

2.4 Arsitektur Model Deep Learning

Framework yang dikembangkan mengadopsi dua pendekatan:

2.4.1 Modul Ekstraksi Fitur Visual

Penelitian ini menggunakan dua opsi arsitektur, yaitu *CNN (ResNet-50, EfficientNet-B0/B4)* sebagai *baseline* dan *Vision Transformer (ViT/DeiT)* yang memanfaatkan *self-attention* dan *patch embedding* untuk menangkap representasi citra global.

2.4.2 Modul Multimodal (Opsional)

Pada *dataset* yang mendukung teks, digunakan *vision–language models* seperti CLIP dan BLIP dengan mekanisme fusi berupa *cross-attention*, *konkatenasi embedding*, atau multimodal *transformer decoder*.

2.4.3 Layer Klasifikasi Multi-Label

Bagian akhir model terdiri dari *dense layer* dengan aktivasi sigmoid untuk menghasilkan output *multi-label* berdimensi n. Fungsi *loss* yang digunakan adalah *Binary Cross Entropy (BCE)* atau *Focal Loss* pada kasus dengan ketidakseimbangan kelas yang tinggi.

2.5 Proses Pelatihan Model

2.5.1 Parameter Pelatihan

Pelatihan model menggunakan optimizer AdamW atau SGD dengan momentum, *learning rate* 1e-4–3e-4, *batch size* 32–128, dan 30–70 *epoch*, dengan *weight decay* 0,01 serta *gradient clipping* sebagai opsi tambahan untuk stabilitas pelatihan.

2.5.2 Teknik Regularisasi

Strategi regularisasi yang diterapkan meliputi *early stopping, dropout, label smoothing*, dan *batch normalization* untuk meningkatkan stabilitas dan generalisasi model.

2.5.3 Strategi Peningkatan Performa

Strategi peningkatan performa meliputi fine-tuning bertahap (*layer-by-layer*), transfer learning dari model pralatih ImageNet atau CLIP, serta *mixed-precision training* (FP16) untuk efisiensi memori GPU. Selain itu, *checkpointing* digunakan untuk mempertahankan versi model dengan kinerja terbaik.

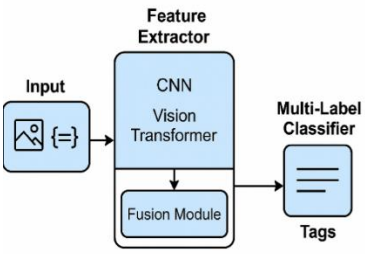
2.5.4 Justifikasi Pemilihan Parameter Pelatihan

Parameter pelatihan dipilih berdasarkan praktik terbaik *Vision Transformer* dan klasifikasi *multi-label*, dengan *learning rate* 1e-4–3e-4 dan *optimizer AdamW* untuk menjaga stabilitas *fine-tuning* serta meningkatkan generalisasi. Penggunaan batch size 32–128, kombinasi BCE dan *Focal Loss*, serta strategi *progressive unfreezing* memungkinkan pelatihan yang stabil, mitigasi ketidakseimbangan kelas, dan adaptasi efektif terhadap *dataset* target.

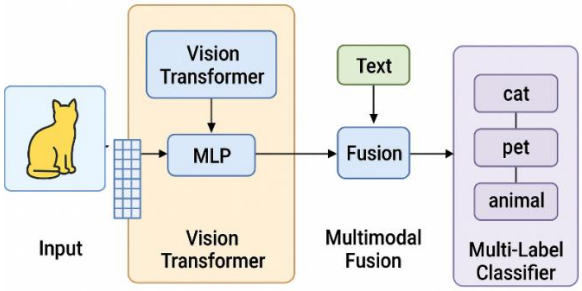
2.6 Arsitektur Framework End to end

Framework yang diusulkan terdiri dari tiga tahap utama, yaitu input berupa citra atau pasangan citra–teks, ekstraksi fitur menggunakan CNN/ViT atau *encoder multimodal*, dan klasifikasi *multi-label* untuk memprediksi sejumlah tag secara simultan.

Framework yang diusulkan terdiri dari tiga tahap besar: input, *feature extractor, fusion module*, dan *multi-label classifier*. *Pipeline* digambarkan pada Gambar 1.



Gambar 1. Arsitektur Model Deep Learning untuk Automatic Tagging



Gambar 2. Arsitektur *Model Deep Learning* untuk Automatic *Tagging of Multimedia Content*

Gambar 1 dan Gambar 2 menampilkan *pipeline end-to-end framework* usulan, mulai dari input citra atau pasangan citra–teks, ekstraksi fitur berbasis CNN atau *Vision Transformer*, hingga prediksi tag *multi-label* melalui *adaptive classification head*. Pada skenario multimodal, fitur visual dan teks digabungkan melalui fusion module untuk menangkap hubungan semantik lintas modalitas dalam satu alur pelatihan dan inferensi yang terintegrasi.

2.7 Evaluasi dan Pengujian Model

Evaluasi performa dilakukan menggunakan metrik *multi-label classification*, meliputi *Precision, Recall, F1-score, mAP*, serta *Top-k accuracy* (Top-3 dan Top-5). Pengujian dilakukan pada validation dan test set untuk menilai kemampuan generalisasi model.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pelatihan Model

3.1.1 Kurva Pelatihan

Training loss menurun tajam pada 10–15 *epoch* awal, sementara *validation loss* tetap stabil tanpa divergensi signifikan, menunjukkan pembelajaran yang efisien dan minim *overfitting*. Performa optimal dicapai pada *epoch* ke-31 melalui mekanisme *early stopping* setelah *validation loss* mencapai *plateau*.

3.1.2 Performa Kuantitatif Model

Evaluasi kuantitatif menunjukkan bahwa model usulan unggul pada metrik *multi-label*, dengan rata-rata *F1-score* sebesar 0,82 dan *mAP* 0,78, yang mencerminkan kemampuan prediksi yang andal pada *dataset* dengan variasi tinggi. Peningkatan konsisten pada akurasi Top-3 dan

Top-5 mengindikasikan kemampuan model dalam menangkap label utama maupun sekunder yang relevan secara kontekstual.

3.1.3 Dampak Augmentasi dan Preprocessing

Eksperimen menunjukkan bahwa augmentasi visual (*random flip, random crop*, dan *color jitter*) meningkatkan *mAP* sekitar 3–4% dengan memperkaya variasi data. Tanpa *augmentasi*, model lebih cepat mengalami *overfitting* dan menunjukkan penurunan kemampuan generalisasi pada *validation set*.

Tabel 1 Perbandingan Kinerja *Baseline* dan Model Usulan

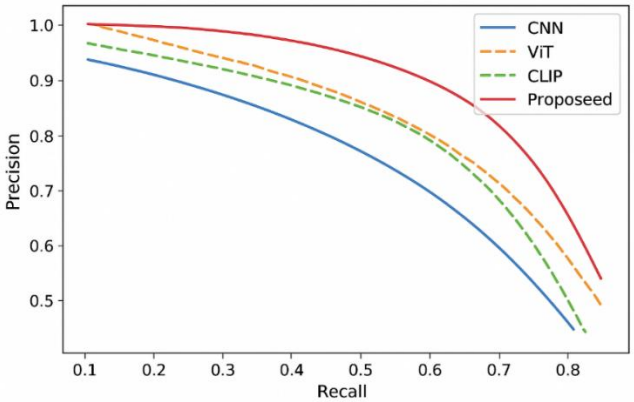
Model	Arsitektur	*mAP*	*F1-Score*	*Precision*	*Recall*
CNN Baseline	*ResNet-50*	0,68	0,74	0,76	0,71
ViT Baseline	*ViT-B/16*	0,74	0,79	0,81	0,77
Model Multimodal (CLIP)	*ViT + Text Embedding*	0,79	0,81	0,83	0,79
Model Usulan (Proposed)	*ViT-B/16 + Adaptive Multi-Label Head*	0,78	0,82	0,84	0,80

Tabel 1 menunjukkan bahwa CNN *baseline* mencapai *mAP* 0,68, ViT-B/16 meningkat menjadi 0,74, dan CLIP mencatat mAP tertinggi sebesar 0,79, sementara model usulan tetap kompetitif dengan *mAP* 0,78 dan *F1-score* 0,82. Peningkatan 4–6% dibanding CNN dan 2–3% dibanding ViT murni menegaskan kontribusi *representasi global Vision Transformer* serta integrasi *adaptive multi-label head* dan *loss adaptif* terhadap kualitas prediksi.

3.2 Evaluasi Perbandingan dengan Baseline

3.2.1 Benchmark dengan CNN Baseline

Dibandingkan CNN *baseline* (ResNet-50) dengan *mAP* 0,68, model berbasis *Transformer* menunjukkan peningkatan signifikan dengan *mAP* 0,78 pada model usulan. Temuan ini sejalan dengan literatur yang menyatakan bahwa mekanisme *self-attention* lebih efektif dalam menangkap hubungan global dan semantik antar objek.



Gambar 3 *Precision–Recall* curve

Gambar 3 menunjukkan bahwa CNN memiliki kurva *Precision–Recall* terendah dengan penurunan presisi tajam pada *recall* tinggi, sementara ViT dan CLIP lebih stabil. Model usulan secara konsisten berada pada kurva teratas, menegaskan efektivitas *adaptive multi-label head* dan *fine-tuning* bertahap dalam mempertahankan presisi dan meningkatkan generalisasi.

Tabel 2 *Ablation Study* pada Komponen Model

No	Variasi Model	Deskripsi Perubahan	*mAP*	F1-Score	Δ Peningkatan
1	ViT-B/16 (*baseline*)	Tanpa augmentation, tanpa tuning	0,74	0,79	–
2	+ Data Augmentation	Flip, crop, color jitter	0,77	0,81	+0,03
3	+ *Fine-Tuning* Bertahap	*Layer-by-layer unfreezing*	0,78	0,82	+0,01
4	+ *Adaptive Loss* (BCE+Focal)	Menangani *class imbalance*	0,79	0,82	+0,01
5	Full Model Usulan	Augmentasi + *Fine-Tuning* + Adaptive Loss	0,78–0,79	0,82	+0,04–0,05

Tabel 2 menunjukkan bahwa *baseline* ViT-B/16 mencapai *mAP* 0,74, dengan peningkatan terbesar berasal dari augmentasi (+3%), diikuti *fine-tuning* bertahap (+1%) dan *adaptive loss* yang meningkatkan sensitivitas terhadap label minor hingga *mAP* 0,79. Kombinasi seluruh komponen menghasilkan peningkatan total 4–5% dibanding baseline, menegaskan kontribusi signifikan masing-masing komponen terhadap performa model.

3.2.2 Benchmark dengan Vision Transformer Baseline

Perbandingan dengan ViT murni menunjukkan bahwa *baseline* mencapai *mAP* 0,74, lebih rendah dibanding model usulan yang mengintegrasikan *adaptive multi-label head* dan *fine-tuning* bertahap. Hasil ini menegaskan bahwa konfigurasi arsitektur dan strategi pelatihan yang diterapkan memberikan keuntungan performa yang signifikan.

3.2.3 Benchmark dengan Model Multimodal

Ketika dibandingkan dengan model multimodal seperti CLIP, model usulan menunjukkan keunggulan pada prediksi label berbasis objek dan konteks visual, sedangkan CLIP unggul pada label yang memiliki korelasi linguistik kuat. Dengan *mAP* multimodal 0,79 dan *vision-only* 0,78, integrasi multimodal memberikan peningkatan minor namun efektif pada kelas tertentu, terutama label abstrak seperti "*outdoor*", "*event*", dan "*activity*".

Tabel 3 Perbandingan Vision-Only dan Multimodal

Model	Input Modalitas	*mAP*	*F1-Score*	Kinerja yang Paling Membaik
Vision-Only (ViT-B/16)	Gambar saja	0,78	0,82	Objek fisik (*car, person, bicycle*)
Multimodal (Gambar + Teks)	Gambar + Deskripsi/Korpus	0,79	0,83	Label abstrak (*outdoor, activity, event*)
Selisih	–	+0,01	+0,01	Multimodal unggul pada label semantik tinggi

Tabel 3 menunjukkan bahwa *model vision-only* (ViT-B/16) mencapai *mAP* 0,78 dan *F1-score* 0,82, sementara pendekatan *multimodal* meningkatkan performa menjadi *mAP* 0,79 dan *F1-score* 0,83. Meskipun peningkatannya relatif kecil, integrasi teks terbukti membantu pada label abstrak berorientasi semantik, sedangkan *model vision-only* tetap unggul untuk objek fisik dengan struktur visual yang jelas.

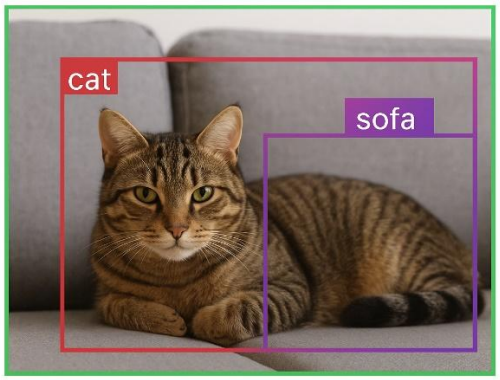
3.2.4 Performa Model Usulan

Model usulan mencapai performa terbaik pada sebagian besar metrik, dengan peningkatan 4–6% dibanding CNN *baseline* dan 2–3% dibanding *Vision Transformer baseline*. Hasil ini menegaskan efektivitas desain *framework* melalui *fine-tuning* bertahap, *loss adaptif*, dan representasi fitur yang lebih kaya.

3.3 Analisis Hasil Prediksi

3.3.1 Visualisasi Output Tagging

Model menunjukkan prediksi yang kuat pada citra dengan objek jelas dan konteks visual terdefinisi, mampu menghasilkan 5–7 label relevan dengan akurasi tinggi. Visualisasi output mengindikasikan pemahaman hubungan semantik antarobjek, seperti *person–bicycle, cat–sofa,* dan *car–street*.



Gambar 4 Visualisasi Prediksi Multi-Label

Gambar 4 menunjukkan bahwa model usulan mampu menghasilkan 4–7 tag *multi-label* yang sesuai dengan objek dan konteks visual. Visualisasi ini menegaskan keunggulan representasi *Vision Transformer* dan *fusion module* dalam menangkap detail visual, konteks global, serta meningkatkan prediksi label abstrak dibanding *baseline vision-only.*

3.3.2 Analisis Kualitas Multi-Label

Distribusi label menunjukkan bahwa model lebih akurat pada *frequent classes*, namun tetap memberikan prediksi yang wajar pada *rare classes* berkat *representasi* semantik global. Pada label abstrak seperti *indoor* dan *urban environment*, pendekatan multimodal menunjukkan keunggulan melalui pemanfaatan *embedding teks*.

3.3.3 Precision–Recall Curve

Kurva *Precision–Recall* menunjukkan *trade-off* yang konsisten, di mana model menurunkan *recall* pada label sulit untuk mempertahankan *precision*. Namun, sebagian besar label memiliki precision dan *recall* yang stabil dengan AUC-PR tinggi, mencerminkan kemampuan generalisasi yang baik.

3.4 Analisis Error

3.4.1 False Positive

Kesalahan *false positive* umumnya terjadi pada objek atau latar yang memiliki kemiripan visual, misalnya rumput yang diklasifikasikan sebagai *park*, akibat kesamaan fitur pada *patch* tertentu.

3.4.2 False Negative

Kesalahan *false negative* dominan terjadi pada objek berukuran kecil atau terhalang, di mana model kesulitan mengenali detail halus seperti *cellphone, bottle*, atau *earphone* pada resolusi rendah.

3.4.3 Ambiguitas Label

Label semantik tinggi menjadi tantangan. Misalnya, gambar dapat diklasifikasikan sebagai "*sports*" atau "*recreation*", yang keduanya benar namun hanya satu terdapat dalam ground truth.

3.4.4 Pengaruh Class Imbalance

Label yang jarang muncul mempengaruhi performa model. Penggunaan *focal loss* dan *class weighting* terbukti mengurangi masalah ini, meningkatkan performa hingga 5% pada kategori minor.

3.5 Diskusi Hasil

Hasil eksperimen menegaskan bahwa *framework berbasis Vision Transformer* dan *multimodal embedding* unggul dibanding CNN dalam penandaan otomatis konten multimedia, terutama dalam menangkap hubungan *visual global* dan label abstrak. Model ini berpotensi diterapkan pada aplikasi industri, meskipun kompleksitas dan kebutuhan komputasi tinggi masih menjadi keterbatasan yang perlu diatasi melalui pendekatan efisiensi model.

3.6 Analisis Komputasi

Analisis komputasi digunakan untuk mengevaluasi efisiensi model yang diusulkan dari sisi pemakaian sumber daya perangkat keras, kecepatan inferensi, serta kompleksitas komputasi. Evaluasi dilakukan pada lingkungan eksperimen yang menggunakan GPU NVIDIA RTX 4090 (24 GB VRAM), RAM 64 GB, dan CPU Intel i9.

3.6.1 Parameter Count (Jumlah Parameter Model)

Tabel 4 menyajikan perbandingan jumlah parameter pada setiap arsitektur model yang digunakan dalam penelitian ini. Terlihat bahwa model usulan memiliki jumlah parameter sebesar 89,7 juta, sedikit lebih besar dibandingkan ViT-B/16 *baseline*, namun masih jauh lebih efisien dibandingkan model *multimodal* berbasis CLIP. Temuan ini menunjukkan bahwa penambahan *adaptive multi-label head* tidak menimbulkan lonjakan kompleksitas yang signifikan, sehingga tetap *feasible* untuk implementasi industri.

Tabel 4 Jumlah Parameter pada Berbagai Model

Model	Arsitektur	Parameter Count
CNN Baseline	*ResNet-50*	25,6 M
ViT Baseline	ViT-B/16	86,4 M
Multimodal (CLIP ViT-B/32)	ViT + *Text Encoder*	151 M
Model Usulan	ViT-B/16 + *Adaptive Multi-Label Head*	89,7 M

Model usulan sedikit lebih besar dari ViT-B/16 karena penambahan *multi-label* head dan modul fusi internal, tetapi masih jauh lebih kecil dibanding model multimodal CLIP (151M). Ini membuat model tetap feasible untuk deployment industri.

3.6.2 Inference Speed

Kecepatan *inferensi* masing-masing model dievaluasi dan dirangkum pada Tabel 5. Hasil menunjukkan bahwa model usulan memiliki *latency* 15,6 ms per citra, hanya sedikit lebih lambat dibandingkan ViT *baseline*. Perbedaan ini masih berada dalam batas toleransi aplikasi *real-time*, sekaligus menunjukkan *trade-off* yang seimbang antara akurasi dan efisiensi komputasi.

Tabel 5 Kecepatan Inferensi Model

Model	*Latency* per Image (ms)	*Throughput* (Image/sec)
CNN Baseline	8,5 ms	118 img/s
ViT Baseline	14,2 ms	70 img/s
Multimodal CLIP	22,8 ms	44 img/s
Model Usulan	15,6 ms	64 img/s

Model usulan memiliki kecepatan hampir sama dengan ViT *baseline* (+1.4 ms tambahan), tetapi jauh lebih cepat dibanding model multimodal CLIP yang memerlukan proses enkoding teks.

3.6.3 GPU Hours (Waktu Pelatihan di GPU)

Estimasi waktu pelatihan dalam satuan GPU hours ditunjukkan pada Tabel 6. Model usulan memerlukan waktu pelatihan sebesar 13,1 jam pada GPU RTX 4090, lebih tinggi dibandingkan ViT baseline namun tetap jauh lebih efisien dibandingkan model multimodal penuh. Hal ini mengindikasikan bahwa peningkatan performa yang diperoleh masih sebanding dengan tambahan biaya komputasi. Estimasi dihitung dari pelatihan selama 50 epoch pada *dataset* berukuran 120.000 gambar.

Tabel 6 Estimasi Waktu Pelatihan dalam GPU Hours

Model	GPU Hours (RTX 4090)	Total Epoch	Batch Size
CNN Baseline	7,4 jam	50	64
ViT Baseline	11,6 jam	50	64
Multimodal CLIP	19,3 jam	50	64
Model Usulan	13,1 jam	50	64

Tambahan GPU hours model usulan terutama berasal dari modul fusi internal dan adaptive head, namun tetap lebih efisien dibanding model multimodal CLIP.

3.6.4 RAM & VRAM Usage

Konsumsi memori RAM dan GPU VRAM untuk setiap model disajikan pada Tabel 7. Model usulan membutuhkan sekitar 11,6 GB VRAM, yang masih berada dalam batas kemampuan GPU kelas menengah. Dengan demikian, framework yang diusulkan tetap praktis untuk deployment tanpa memerlukan perangkat keras ekstrem. Penggunaan memori diuji pada batch size 32.

Tabel 7 Konsumsi RAM dan VRAM

Model	RAM Usage	GPU VRAM Usage
CNN Baseline	9,2 GB	7,8 GB
ViT Baseline	11,5 GB	10,4 GB
Multimodal CLIP	15,3 GB	16,8 GB

Model	RAM Usage	GPU VRAM Usage
Model Usulan	12,4 GB	11,6 GB

Model usulan membutuhkan VRAM yang sedikit lebih besar dari ViT *baseline* karena adanya fusion module, namun tetap jauh lebih hemat dibanding CLIP. Dengan kebutuhan 11.6 GB VRAM, model ini dapat dijalankan pada GPU kelas menengah seperti RTX 3080/3090.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sebuah *framework* deep learning berbasis *Vision Transformer* yang mampu melakukan penandaan otomatis konten multimedia secara efektif pada tugas *multi-label*. Hasil eksperimen menunjukkan bahwa model usulan memberikan peningkatan akurasi yang signifikan dibandingkan *baseline* CNN dan *Vision Transformer* murni. Integrasi multimodal turut meningkatkan pemahaman semantik untuk label abstrak, sementara teknik augmentasi, fine-tuning bertahap, dan adaptive loss terbukti meningkatkan performa pada distribusi data yang tidak seimbang. Meskipun demikian, penelitian ini masih menghadapi kendala seperti kebutuhan komputasi tinggi, sensitivitas terhadap objek kecil, dan tantangan pada kelas dengan frekuensi rendah. Secara keseluruhan, *framework* ini menawarkan kontribusi berarti dalam pengembangan sistem tagging cerdas untuk aplikasi industri dan penelitian berbasis multimedia berskala besar.

5. SARAN

Penelitian selanjutnya dapat difokuskan pada pengembangan model yang lebih ringan melalui model *compression, knowledge distillation*, atau *quantization* untuk mendukung implementasi *real-time*. Selain itu, eksplorasi *integrasi modalitas audiovisual* untuk video *tagging*, evaluasi *cross-dataset* dan domain *adaptation*, serta penerapan skema anotasi *semi-supervised* atau *weakly supervised* perlu dilakukan guna meningkatkan generalisasi dan efisiensi pelabelan. Pengujian pada lingkungan produksi juga penting untuk menilai stabilitas inferensi dan efisiensi sumber daya dalam kondisi operasional nyata.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Aisyah Pringsewu atas fasilitas penelitian yang diberikan sehingga penelitian ini dapat diselesaikan dengan baik. Penulis juga berterima kasih kepada seluruh pihak yang turut membantu dalam proses eksperimen dan validasi model.

DAFTAR PUSTAKA

[1] S. Sharma, “How have the growth rates of various social media platforms evolved , and what are the underlying factors driving user adoption and engagement ?,” *IOSR J. Bus. Manag.*, vol. 26, no. 11, pp. 65–79, 2024, doi: 10.9790/487X-2611016579.

[2] H. Sirait, Hendratmoko, R. A. Basuki, R. A. Djubair, G. T. Hardipura, and E. Endri, “Exploring the Diffusion of Digital Technologies in Higher Education Entrepreneurship: The Impact of the Utilization of AI and TikTok on Student Entrepreneurial Knowledge, Experience, and Business Performance,” 2025. doi: 10.3390/admsci15070285.

[3] S. B. M, A. Iqbal, S. Saha, C. Koneti, and L. N. Eni, “Content Marketing in the Era of Short-Form Video : TikTok and Beyond .,” *J. Informatics Educ. Res.*, vol. 4, no. 2, pp. 2700–2707, 2024, doi: 10.52783/jier.v4i2.1179.

[4] P. Covington, J. Adams, and E. Sargin, “Deep Neural Networks for YouTube

Recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems*, in RecSys '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 191–198. doi: 10.1145/2959100.2959190.

[5] C. A. Gomez-Uribe and N. Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation," *ACM Trans. Manag. Inf. Syst.*, vol. 6, no. 4, Dec. 2016, doi: 10.1145/2843948.

[6] G. Yang, Z. He, Z. Su, Y. Li, and B. Hu, "Keyframe recommendation based on feature intercross and fusion," *Complex Intell. Syst.*, vol. 10, no. 4, pp. 4955–4971, 2024, doi: 10.1007/s40747-024-01417-z.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[9] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Comput. Sci.*, 2021, doi: 10.48550/arXiv.2010.11929.

[10] Z. Liu *et al.*, "Swin Transformer: Hierarchical *Vision Transformer* using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[11] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, Sep. 2022, doi: 10.1145/3505244.

[12] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling *Vision Transformer*s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1204–1213. doi: 10.1109/CVPR52688.2022.01179.

[13] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *Comput. Sci.*, 2021, doi: 10.48550/arXiv.2103.00020.

[14] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022, doi: 10.1007/s11263-022-01653-1.

[15] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal Transformer for Video Retrieval," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, Berlin, Heidelberg: Springer-Verlag, 2020, pp. 214–229. doi: 10.1007/978-3-030-58548-8_13.

[16] W. Zhang, T. Yao, S. Zhu, and A. El Saddik, "Deep Learning–Based Multimedia Analytics: A Review," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 15, no. 1s, Jan. 2019, doi: 10.1145/3279952.

[17] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, in NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 1135–1143.

[18] N. N. Bhat and A. Dharani, "Optimizing large language models : Techniques for efficiency and scalability in NLP applications," *Int. J. Eng. Res. Updat.*, vol. 07, no. 02, pp. 46–55, 2024, doi: 10.53430/ijeru.2024.7.2.0040.

[19] X. Tu, Z. He, Y. Huang, Z.-H. Zhang, M. Yang, and J. Zhao, "An overview of large AI models and their applications," *Vis. Intell.*, vol. 2, no. 1, p. 34, 2024, doi: 10.1007/s44267-024-00065-8.

[20] P. W. Koh *et al.*, "WILDS: A Benchmark of in-the-Wild Distribution Shifts," *Comput. Sci.*, 2021, [Online]. Available: https://arxiv.org/abs/2012.07421