Progresif: Jurnal Ilmiah Komputer
https://ojs.stmik-banjarbaru.ac.id/index.php/progresif/index
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat - Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877

Analisis Komparatif Unjuk Kerja Model *Vision Transformers* Dengan *ConvNeXt* Dalam Rekognisi Citra Warangka Keris Bali

DOI: http://dx.doi.org/10.35889/progresif.v21i2.2987

Ni Putu Widya Yuniari[1*], Gde Wikan Pradnya Dana[2], I Gede Wira Darma[3]
Teknik Komputer, Universitas Warmadewa, Denpasar, Indonesia
*e-mail *Corresponding Author:* putu.widyayuniari@warmadewa.ac.id

Abstract

The application of attention mechanisms in image recognition has emerged as a new paradigm in computer vision, serving as a foundational approach in generative AI. Two state-of-the-art models frequently referenced in recent studies are Vision Transformers (ViT), introduced by Google, and ConvNeXt, developed by Meta (Facebook) AI Research. However, their application in recognizing local cultural imagery, such as the warangka (sheath) of the Balinese keris, remains highly limited. The urgency of this study lies in evaluating the effectiveness of AI models in supporting technology-based cultural preservation. This study aims to compare the unjuk kerjance of these two models in handling the classification and recognition of warangka keris (Balinese kris sheaths). The methodology involves data augmentation, feature extraction, patch processing (for ViT), model construction, evaluation, and image recognition analysis using Grad-CAM. The dataset comprises a combination of primary and secondary sources. Primary data were collected through field visits to kris-making workshops in Bali, while secondary data were obtained from previous studies. The kris sheath image classes used in this study include: 'Sesrengatan', 'Kojongan', 'Batun Poh', 'Kekandikan', and 'Beblatungan'. The study successfully developed image classification models, achieving an accuracy of 82% with the ViT model and 97% with the ConvNeXt model. The recognition process effectively highlighted the most significant regions of each image, providing valuable insight for future generative AI research.

Keywords: *Attention, ConvNeXt, Keris Bali, Vision Transformers*

Abstrak

Penerapan *attention* dalam rekognisi citra menjadi pendekatan baru dalam pengenalan gambar dan berpotensi menjadi benchmark dalam pengembangan kecerdasan buatan generatif. Dua model terkini yang banyak diteliti adalah *Vision Transformers* (ViT) dari Google dan *ConvNeXt* dari Meta AI. Namun, penerapan keduanya dalam pengenalan citra budaya lokal seperti warangka keris Bali masih sangat terbatas. Urgensi penelitian ini terletak pada upaya mengevaluasi efektivitas model kecerdasan buatan dalam mendukung pelestarian budaya berbasis teknologi. Penelitian ini bertujuan untuk membandingkan performa ViT dan *ConvNeXt* dalam klasifikasi serta rekognisi citra warangka keris Bali. Metode yang digunakan meliputi augmentasi data, ekstraksi fitur, proses patching (untuk ViT), pembuatan model, pengujian, serta analisis grad cam. Data yang digunakan merupakan gabungan data primer (hasil kunjungan ke workshop pembuatan keris Bali) dan data sekunder dari berbagai sumber. Citra keris yang digunakan antara lain: 'Sesrengatan', 'Kojongan', 'Batun Poh', 'Kekandikan', dan 'Beblatungan'. Hasil menunjukkan akurasi 82% (ViT) dan 97% (ConvNeXt), serta bagian penting citra berhasil dikenali sebagai benchmark generatif.

Kata kunci: *Attention; ConvNeXt; Keris Bali; Vision Transformers*

1. Pendahuluan

.Dalam beberapa tahun terakhir, bidang *computer vision* (pengenalan gambar oleh komputer) mengalami kemajuan pesat, terutama sejak diperkenalkannya mekanisme *attention* dalam tugas-tugas rekognisi dan pengenalan gambar oleh kecerdasan buatan (AI) [1]. Mekanisme ini memungkinkan komputer fokus pada bagian terpenting dari sebuah gambar, baik dengan maupun tanpa teknik konvolusi. Kemajuan ini dimulai sejak diterbitkannya paper “Attention is All You Need” oleh Google Brain dan University of Toronto pada tahun 2017 [2]. Awalnya digunakan di bidang *Natural Language Processing* (NLP), model-model ini kini telah dikembangkan untuk tugas-tugas visual. Teknologi ini juga menjadi dasar penting bagi *generative AI* (kecerdasan buatan generative) yang bersifat multimodal, mampu menggabungkan teks dan gambar dalam satu sistem [3]. Dalam konteks pelestarian budaya, pendekatan ini penting untuk merekam dan memperkenalkan objek budaya yang bersifat benda (*tangible*) secara digital [4]. Teknik ini juga memungkinkan dokumentasi budaya dilakukan dalam bentuk digital yang dapat dipertanggungjawabkan keasliannya secara objektif [5], dan pada akhirnya bisa menjadi sarana baru dalam pelestarian serta alih media objek-objek budaya [6].

Salah satu objek budaya yang penting untuk dilestarikan adalah keris Bali. Keris bukan sekadar perlengkapan upacara, tetapi menyimpan makna filosofis, historis, dan menjadi bagian penting dari identitas sosial masyarakat Bali [7][8]. Dalam antropologi Bali, keris juga menyimpan cerita, ritus, hingga mantra yang bersifat multimodal [9]. Bahkan, keris dapat mewakili eksistensi pemiliknya secara simbolik maupun langsung [10][11]. Urgensi ini semakin nyata karena minat generasi muda terhadap profesi *pande* (pembuat keris) terus menurun. Salah-satunya terjadi pada Desa Sawan dimana jumlah wangsa pande menurun dari 45 orang di tahun 2021 hingga hanya menjadi 43 orang di tahun 2023[12]. Bahkan di beberapa desa seperti Bukti, Bengkala, dan Sanggalangit, tidak ditemukan lagi pembuat keris [13][14][15]. Fakta ini menunjukkan bahwa pelestarian keris tidak bisa lagi hanya mengandalkan pewarisan tradisional. Diperlukan media dan pendekatan baru yang mampu mendokumentasikan serta menyimpan nilai budaya keris secara digital agar tidak hilang ditelan waktu [16]. Salah satu penelitian terbaru yang menerapkan mekanisme *attention* dalam *computer vision* khususnya pada pendekatan preservasi budaya adalah penelitian yang dilakukan oleh Sihananto dkk pada tahun 2024. Penelitian ini membandingkan unjuk kerja empat model *deep learning* dalam klasifikasi jenis-jenis Wayang. Adapun model yang diuji adalah satu model berbasis *transformers* yaitu *Vision Transformers* (ViT) dan tiga buah model berbasis konvolusi yaitu ResNet, YOLOv5 dan YOLOv8. Penelitian ini menunjukan bahwa model *Vision Transformers* memiliki kinerja yang paling barik dengan akurasi mencapai 91,3%. Penelitian ini menyimpulkan bahwa mekanisme *attention* seperti ViT mampu menangkap kompleksitas visual Wayang dengan sangat baik [17].

Penelitian kedua dilakukan oleh Tran dkk pada tahun 2025. Penelitian ini membahas klasifikasi citra *Intangible Cultural Heritage* (warisan budaya takbenda) di kawasan Delta sungai Mekong, Vietnam. Penelitian ini mengklasifikasikan gambar ICH ke dalam 17 kategori, yang mencakup berbagai praktik tradisional, pertunjukan, dan ekspresi budaya khas Delta Mekong. Uniknya penelitian ini hanya menggunakan *Vision Trasnformers* (ViT) untuk ekstraksi fitur dan kemudian dilakukan mekanisme *fine tuning* untuk membuat model baru degan konsep *stacking* menggunakan *logistic regression*. Penelitian ini menyimpulkan bahwa model *Vision Transformer* (ViT) memiliki potensi kuat dalam menangkap kompleksitas visual ICH yang dikenal rumit [18]. Penelitian ketiga dilakukan oleh Pei dkk pada tahun 2023. Penelitian ini berfokus pada pengenalan material benda peninggalan budaya menggunakan *computer vision* dan *attention mechanism*, dengan tujuan memahami hubungan antara bahan benda budaya dan atribut budaya dari dinasti tertentu. dalam hal ini, budaya Tiongkok tradisional. Penelitian dilakukan dengan memperkaya model berbasis konvolusi yaitu *Efficient Net* dengan menggunakan mekanisme *attention*. Teknik ini diperlukan untuk menekankan bagian gambar yang paling relevan untuk pengenalan material. Penelitian ini menghasilkan akurasi pengenalan material mencapai 88,15%, dengan rata-rata presisi sebesar 83,3% [19].

Penelitian keempat dengan pendekatan multimodal dilakukan oleh Fan dkk pada tahun 2023. Penelitian ini berfokus pada pengembangan model klasifikasi citra warisan budaya takbenda pada lukisan tradisional untuk Tahun Baru Imlek dan patung-patung dari tanah liat. Penelitian ini bertujuan membantu masyarakat mengenali serta melestarikan budaya tersebut. Penelitian ini menerapkan mekanisme *attention* yang berfokus pada fitur visual dari gambar, dengan mempertimbangkan deskripsi teks yang menyertai gambar. Model yang digunakan adalah *Multimodal Interaction and Cross-Modal Learning Framework* (MICMLF) yang menggabungkan *Multimodal Attention* dan *Hierarchical Fusion*. Penelitian ini mengungguli beberapa metode canggih lainnya dalam hal akurasi klasifikasi [20]. Penelitian kelima dengan pendekatan *information retrieval* dilakukan oleh Gao dkk pada tahun 2023. Penelitian ini mengembangkan sebuah pendekatan berbasis *deep learning* yang diperkaya dengan mekanisme *attention* untuk klasifikasi dan *image retrieval* (pencarian citra) pada warisan arsitektur diaspora Tionghoa di Jiangmen, Guangdong, Tiongkok. Model yang digunakan dalam penelitian ini adalah *Convolutional Neural Network Attention Retrieval Framework* (CNNAR). Penelitian ini menghasilkan akurasi hingga 98,3% [21].

Berdasarkan penelitian pertama, penelitian kedua, dan penelitian ketiga dapat ditunjukkan bahwa mekanisme *attention* berperan baik dalam proses pengenalan citra kebudayaan yang dikenal rumit seperti: wayang, ukiran, patung, relief dan bahkan arsitektur. Artinya mekanisme ini telah berhasil menjalankan fungsinya untuk menetapkan bagian-bagian terpenting yang membedakan suatu citra dengan citra lainnya, baik dengan mekanisme satu model (penelitian pertama), maupun dengan mekanisme *hybrid* ataupun *stacking* (penelitian kedua dan ketiga). Pada penelitian keempat, kita dapat melihat fungsi lain dari mekanisme *transformers*, yaitu dapat menggabungkan pendekatan multimodal untuk memperkaya informasi citra. Pada penelitian kelima kita dapat melihat implementasinya dimana model-model ViT baik yang sudah diperkaya dengan multimodal maupun tidak, dapat digunakan untuk *information retrieval*. Namun ada dua hal yang perlu dikritisi dan menjadi tantangan dari kelima penelitian diatas. Hal pertama adalah objek kajiannya. Berdasarkan hal ini, ditunjukkan bahwa belum adanya penelitian yang membahas citra senjata tradisional, khususnya Keris Bali dengan pendekatan *attention* secara langsung. Adapun objek kajian senjata tradisional masih berlangsung pada mekanisme konvolusi dengan pendekatan deep learning. Berikutnya, yang menjadi kritik dari berbagai penelitian diatas adalah dominasi model *Vision Transformers* pada berbagai penelitian. Padahal masih terdapat model lain seperti NAS dan *ConvNeXt*. Kendati kurang populer dalam ranah studi akademis, kedua model tersebut dapat memperkaya khazanah pemikiran kita dalam mencari bentuk-bentuk terbaik dalam upaya preservasi budaya berbasis mekanisme attention.

Studi awal ini mengarahkan peneliti untuk mengevaluasi relevansi penggunaan model-model *attention* pada objek budaya berupa senjata tradisional, khususnya Keris Bali, serta mengkaji potensi penerapan model-model alternatif dalam tugas serupa. Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk menganalisis rekognisi citra pada objek senjata tradisional Keris Bali melalui pendekatan komparasi antara model *Vision Transformers* (ViT) dan *ConvNeXt* dengan tujuan mengidentifikasi model yang paling optimal. Penelitian ini diharapkan dapat memperkaya khazanah pengetahuan baik bagi peneliti maupun pengembang, serta menjadi benchmark pengambilan kebijakan dari berbagai stakeholder dalam upaya preservasi budaya berkelanjutan berbasis teknologi.

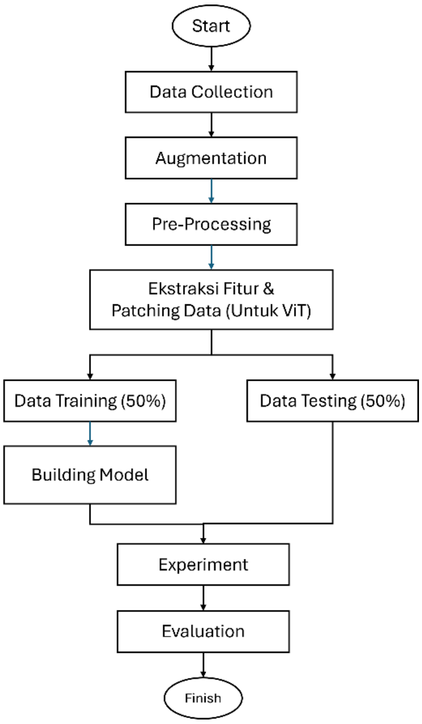
2. Metodologi

1) Alur Penelitian

Adapun alur dari penelitian ini dapat dilihat pada Gambar 1.

Penelitian ini dimulai dengan pengumpulan dataset. Dataset yang digunakan merupakan kombinasi antara data primer dan data sekunder. Data primer didapatkan dengan kunjungan langsung ke workshop pembuatan keris di Bali, sementara data skunder didapatkan dari berbagai penelitian terdahulu. Adapun citra keris yang digunakan adalah: 'Sesrengatan', 'Kojongan', 'Batun Poh', 'Kekandikan', 'Beblatungan'. Penelitian kemudian dilanjutkan dengan melakukan augentasi gambar untuk menghasiLkan berbagai pola gambar yang berbeda. Adapun metode augmentasi yang digunakan adalah: *flip, rotation, noising* dan *contrast engineering*.

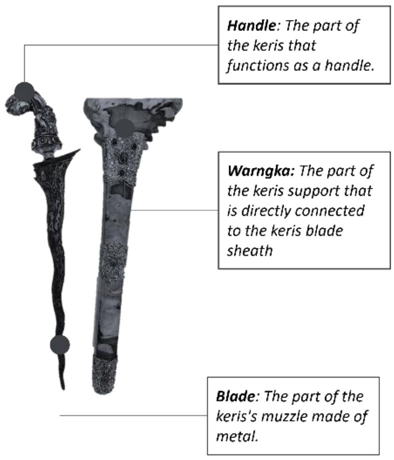
Selanjutnya, penelitian ini dilanjutkan dengan tahap *pre-processing* yang bertujuan untuk menyamakan dimensi gambar menjadi 150px x 150px. Kemudian dilanjutkan dengan proses ekstraksi fitur mengubah data menjadi vector *embedding*. Karena dimensi dari masing-masing gambar sudah di*reengineering* menjadi 150px x 150px, dan vector warna yang digunakan adalah RGB (3 warna), maka dimensi *vector embedding* yang terbentuk dari tiap gambar adalah 150 x 150 x 3. Dimensi ini juga yang akan menjadi dimensi input pada model uji kita. Kemudian kumpulan data vektor tersebut dibagi menjadi *data training* dan *data testing*. *Data training* digunakan untuk pelatihan model, dan data testing digunakan untuk pengujian. Rasio pembagian data training dan data testing adalah 50:50. Rasio ini dipilih untuk mempertahankan kehandalan pada pengujian dengan tetap mempertahankan unjuk kerja pelatihan. HaI ini dikarenakan jumlah data yang tidak imbang dan dikhawatirkan terjadi bias jika jumlah *data training* atau testingnya terlalu sedikit. Kemudian penelitian dilanjutkan dengan membangun dua buah model, yaitu *Vision Transformers* (ViT) dan *ConvNeXt*. Terakhir, untuk memastikan model berjalan dengan sempurna, dilakukan pengujian dengan *Convusion Matrix, Classification Report dan Grad Cam*



Gambar 1. Diagram Alur Penelitian

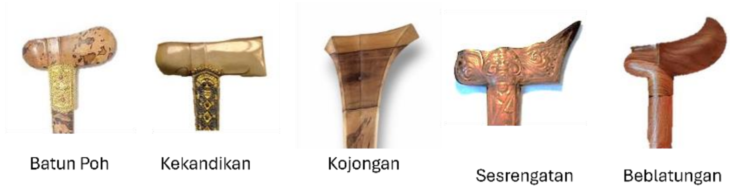
2) Dataset

Dataset yang dikumpulkan pada penelitian ini adalah dataset warangka Keris Bali. Bagian ini dipilih karena menyimpan banyak memori antropologi dan kebudayaan dalam setiap bentuk dan ukirannya. Artinya, bagian ini bukan hanya menyimpan memori estetika, melainkan juga hermeneutika yang membagi masyarakat bali dalam berbagai sub-kultur serta berbagai ritus dan adat kebudayaan lainnya [8]. Bagian ini juga yang paling terlihat daripada bagian keris lainnya, seperti bilah maupun *handle* Keris. Adapun bagian warangka dalam desain Keris Bali dapat dilihat pada Gambar 2.



Gambar 2. Bagian-bagian Keris Bali

Dataset warangka Keris Bali yang digunakan pada penelitian ini, terbagi menjadi data primer dan data skunder. Data primer didapatkan dari pengumpulan data langsung di beberapa workshop pembuatan keris wangsa pande. Sementara data skunder didapatkan dari berbagai penelitian sebelumnya. Adapun jenis warangka yang berhasil dikumpulkan adalah 'Sesrengatan', 'Kojongan', 'Batun Poh', 'Kekandikan' dan 'Beblatungan'. Beberapa sampelnya dapat dilihat pada Gambar 3.



Gambar 3. Sampel Dataset Warangka Keris Bali

3) Augmentasi Gambar

Augmentasi gambar pertujuan untuk melakukan re-engineering gambar pada gambar yang sama untuk menghasilkan pola atau bentuk gambar yang lain [22]. Adapun metode augmentasi yang digunakan pada penelitian ini antara lain: : *flip, rotation, noising & contrast engineering*. Sampelnya dapat dilihat pada Gambar 4.

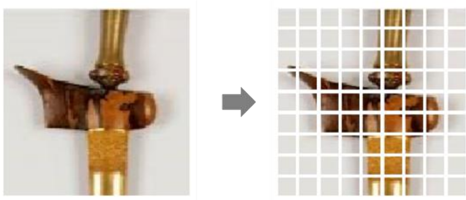
Gambar 4. Sampel Hasil Augmentasi

4) Pre-Processing

Pre-processing merupakan tahapan yang dilakukan dengan memberikan perlakuan khusus pada gambar secara seragam sebelum masuk ke tahap pemrosesan lebih lanjut [23]. Tahapan ini menjadi penting untuk menyamakan dimensi data dan vektor embedding yang terbentuk, sehingga data dapat diproses oleh model. Teknik *pre-processing* yang digunakan dalam penelitian ini adalah menyamakan dimensi data pada dimensi 150px x 150px. Jadi data-data gambar yang memiliki dimensi yang berbeda akan di-kompresimaupun diekspansi pada dimensi tersebut. Teknik ini tidak akan merusak struktur gambar, karena dataset awal yang dimiliki sudah memiliki rasio 1:1, sehingga walaupun dikompresi maupun diekspansi, strukturnya akan tetap sama [24].

5. Patching

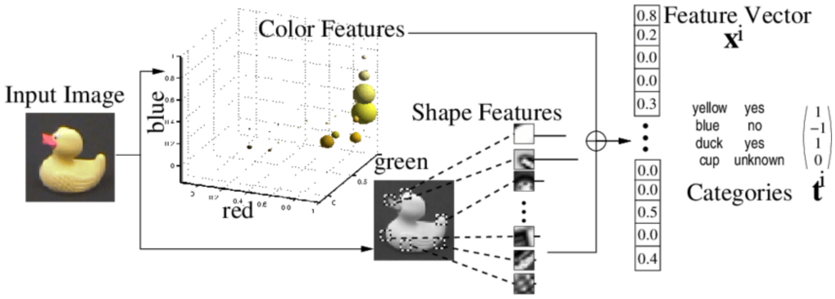
Patching adalah teknik untuk membagi data menjadi kolom-kolom vektor tertentu sebelum memasuki pemorsesan lebih lanjut [25]. Teknik ini diperlukan pada model berbasis Vision Transformers (ViT) sebagai pengganti konvolusi [26]. Hal ini dikarenkan model ViT tidak menggunakan modul konvolusi untuk membagi gambar menjadi patch-patch tertentu dan mengambil bagian terpenting [27]. Salah satu sampel dari proses *patching* pada penelitian ini dapat dilihat pada Gambar 5.



Gambar 5. Sampel Hasil Patching

6) Ekstraksi Fitur

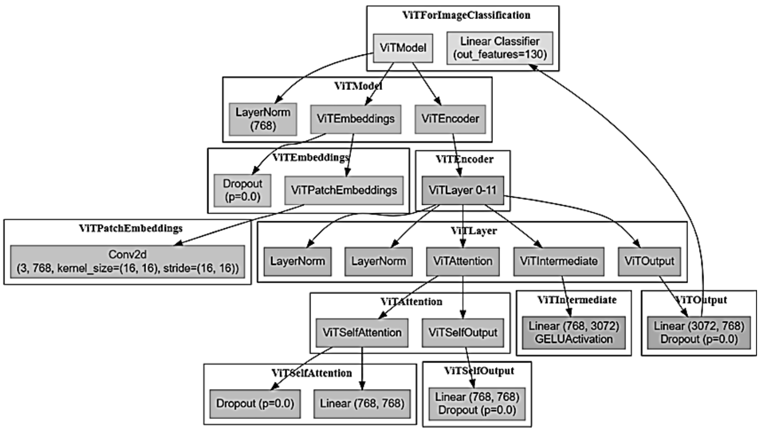
Ekstraksi fitur merupakan proses krusial dalam tugas-tugas pengenalan pola dan analisis citra digital secara komputasi. Tahapan ini bertujuan untuk mereduksi kompleksitas data visual dengan merepresentasikan informasi penting dari sebuah gambar dalam bentuk numerik. Hal ini dikarenakan mesin tidak dapat langsung memproses sebuah gambar, melainkan pola komputasi dari gambar tersebut [28]. Karena dimensi gambar sudah diubah menjadi 150px x 150px dan vektor warna yang digunakan adalah RGB (3 warna), maka vektor hasil ekstraksi fitur pada penelitian ini akan memiliki dimensi 150 x 150 x 3. Skema ekstraksi fitur yang digunakan pada penelitian ini dapat dilihat pada Gambar 6.



Gambar 6. Mekanisme Ekstraksi Fitur Pada Penelitian [28]

7) Arsitektur *Vision Transformers* (ViT)

Vision Transformers (ViT) merupakan sebuah arsitektor *deep learning* berbasis *transformers* yang digunakan untuk tugas-tugas pengolahan citra. Model ini pertama kali diperkenalkan oleh Dosovitskiy et al. pada tahun 2020 yang merupakan tim dari Google Research, khususnya Google Brain [27]. Model ini bertujuan menggantikan arsitektur konvolusional (CNN) tradisional dalam tugas visi komputer [29]. Arsitektur model ViT dapat dilihat pada Gambar 7.

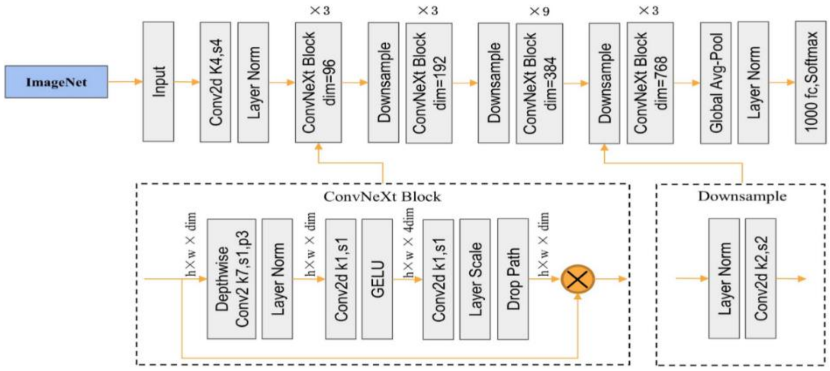


Gambar 7. Arsitektur Vision Transformers (ViT) [29]

Meskipun revolusioner, model ViT juga memiliki tantangan dimana unjuk kerjanya akan sangat turun jika dataset pelatihannya kecil [30]. Untuk mengatasi hal ini, beberapa model ViT kemudian dilakukan proses *fine-tuning* dengan menambahkan beberapa layer konvolusi agar bisa menangkap bagian terpenting dari setiap patch data [31]. Pada penelitian ini, model ViT yang digunakan adalah model hasil *fine-tuning* yang sudah diperkaya dengan arsitektur LeNet menjadi arsitektur LeViT. Model *enhancement* ini dipilih karena ringan dan tidak banyak mengubah struktur model [32].

8) Arsitektur *ConvNeXt*

Arsitektur model *ConvNeXt* dapat dilihat pada Gambar 8.



Gambar 8. Arsitektur Model C*onvNeXt* [34]

ConvNeXt merupakan arsitektur jaringan saraf berbasis konvolusional (CNN) modern yang dikembangkan untuk menjawab tantangan dominasi model *vision transformer* dalam tugas-tugas *computer vision* [33]. Tidak hanya menjadi model alternatif, melainkan melibatkan pengayaan pada beberapa fitur untuk menjawab tantangan ViT pada dataset kecil [34]. Tidak seperti *Vision Transformer* (ViT) yang mengandalkan mekanisme *self-attention*, *ConvNeXt* tetap mempertahankan struktur hierarkis dan efisien namun dengan mengikuti cara berpikir mekanisme *attention* untuk mendapatkan fitur terpenting [35]. ConvNeXt dikembangkan oleh tim Facebook *AI Research* (FAIR) dan pertama kali diperkenalkan pada tahun 2022 [36]. Karena sudah berbasis konvolusi, model ini tidak perlu lagi melakukan patching atau menambahkan layer konvolusi diluar arsitektur utama. Artinya, data hasil ekstraksi fitur dapat langsung digunakan [37].

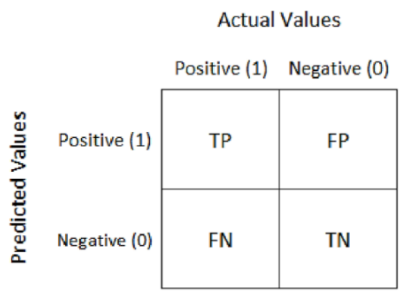
Sama seperti arsitektur *Vision Transformers* (ViT), arsitektur *ConvNeXt* juga sudah diperkaya dengan kemampuan multimodal [38]. Artinya model ini dapat menggabungkan data-data image dengan jenis data lainnya seperti text. Artinya gambar citra warangka Keris Bali yang diperoleh dapat diperkaya dengan keterangan mengenai ritus, teknologi tradisional, hingga objek pemajuan kebudayaan lainnya yang berbasis text. Kemampuan ini menjadikan *ConvNeXt* tetap reliable untuk digunakan pada penelitian ini. Kemampuan ini juga yang membuatnya setara dengan ViT, dan layak untuk dijadikan model alternatif dalam berbagai studi yang melibatkan pemrosesan kontekstual [39].

Arsitektur ConvNeXt memiliki banyak varian, antara lain: *ConvNeXt-T (Tiny), ConvNeXt-S (Small), ConvNeXt-B (Base), ConvNeXt-L (Large), ConvNeXt-XL (Extra Large)* [40]. Semua varian memiliki arsitektur dasar yang sama, hanya jumlah blok dan dimensi channel yang berbeda. Varian ConvNeXt yang digunakan pada penelitian ini adalah *ConvNeXt Tiny*. Hal ini dikarenakan kebutuhan yang relatif kecil yang hanya sebagai benchmark untuk multimodal pada berbagai penelitian berikutnya. Kebutuhan penelitian ini membuatnya tetap valid karena arsitektur ini sering digunakan pada penelitian tahap awal [41]. Varian ini juga terkenal ringan dan efisien, sehingga ocok untuk eksperimen cepat dengan daya komputasi terbatas. Unjuk kerjanya juga tinggi, hampu mengungguli ResNet-50 pada ImageNet walau arsitekturnya lebih kecil [42].

9) Teknik Analisis Data

Analisis dilakukan dengan menguji seluruh data testing pada model ViT dan *ConvNeXt* yang sudah dilatih sebelumnya. Kemudian hasil prediksi tersebut dibandingkan dengan data yang sebenarnya. Dalam melakukan pengujian ini, metode yang digunakan adalah *Confusion Matrix*. *Confusion matrix* adalah matriks pengukuran yang membandingkan antara data hasil prediksi dengan data yang sebenarnya pada setiap kelas [37]. Contoh confusion matrix dapat dilihat pada Gambar 9.

Empat komponen utama dalam *confusion matrix* yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) menjadi dasar untuk menghitung berbagai metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.



Gambar 9. Contoh *Confusion Matrix* [37]

Metrik evaluasi model pertama yang digunakan adalah akurasi. Akurasi mengukur seberapa banyak prediksi model yang benar dibandingkan dengan seluruh prediksi yang dilakukan. Rumus untuk menghitung akurasi model dari confusion matrix adalah sebagai berikut:

$$cc = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

Metrik evaluasi model kedua yang digunakan adalah *precision*. *Precision* mengukur proporsi prediksi positif yang benar. Rumus untuk menghitung precision dari confusion matrix adalah sebagai berikut:

$$recision = \frac{TP}{TP+FP} \tag{2}$$

Metrik evaluasi ketiga yang digunakan adalah *recall*. *Recall* mengukur kemampuan model untuk menemukan semua kasus positif yang sebenarnya ada. Rumus untuk menghitung recall dari *confusion matrix* adalah sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Metrik pengujian terakhir yang digunakan adalah f1-score. F1-Score adalah harmonik rata-rata dari presisi dan recall, digunakan untuk menyeimbangkan keduanya. Rumus untuk menghitung recall dari confusion matrix adalah sebagai berikut:

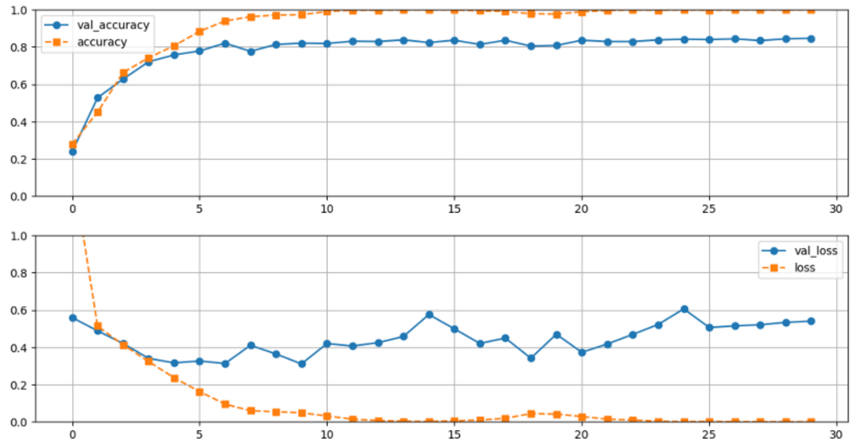
$$F1 = \frac{Precision * Recall}{Precision+Recall} \tag{4}$$

3. Hasil dan Pembahasan

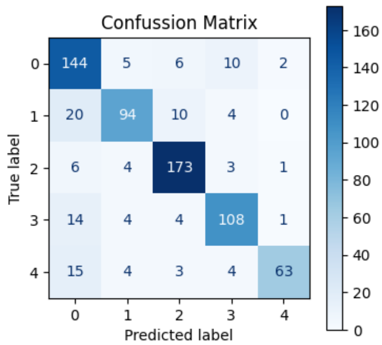
1) Model *Vision Transformers*

Setelah melakukan pelatihan pada 30 iterasi, didapatkan trend akurasi yang menunjukan peningkatan yang signifikan, sementara trend loss menunjukan penurunan yang signifikan. Hal ini mengindikasikan bahwa proses pembelajaran pada model *Vision Transformers* (ViT) berlangsung dengan baik. Grafik unjuk kerja akurasi dan *loss* selama proses *training* pada model *Vision Transformers* (ViT) dapat dilihat pada Gambar 10.

Grafik pada Gambar 10 menunjukan akurasi pelatihan meningkat secara konsisten dari awal hingga akhir proses pelatihan, dengan kecenderungan mencapai nilai mendekati maksimum (sekitar 99%) setelah *epoch* ke-10. Sementara itu, akurasi validasi juga menunjukkan peningkatan yang cukup tajam pada fase awal, lalu mencapai titik stabil di kisaran 82–85% mulai sekitar epoch ke-7. Sementara itu, grafik nilai *loss* memperlihatkan penurunan selama beberapa *epoch* awal dan menunjukan sagnansi setelahnya, yang mengindikasikan keberhasilan optimasi dalam menurunkan fungsi objektif. Stabilitas akurasi dan loss ini menunjukkan bahwa model telah menemukan representasi yang relevan dan dapat digeneralisasi terhadap data yang tidak terlihat sebelumnya.



Gambar 10. Trend Akurasi (atas) dan Loss (bawah)
pada pelatihan model *Vision Transformers* (ViT)



Gambar 11. Confision Matrix Pengujian Model ViT

Setelah proses *training* selesai, kemudian berikutnya akan dilakukan proses pengujian pada model *Vision Transformers* yang sudah dilatih. Tujuannya untuk mengetahui unjuk kerja akhir model pada data pengujian. Proses ini dilakukan dengan menggunakan data testing yang sudah disiapkan. *confusion matrix* dari hasil pengujian model ViT dapat dilihat pada Gambar 11.

Dari Gambar 11 Dapat dilihat bahwa sebagian besar data *testing* berhasil diprediksi dengan baik, hanya saja masih terdapat beberapa error atau kesalahan prediksi yang menghasilkan unjuk kerja yang kurang memuaskan. Kemudian dari *confusion matrix* pada Gambar 11 dapat dihitung akurasi, *precision*, *recall* dan *f1-score* nya. Hasilnya dapat dilihat pada Tabel 1.

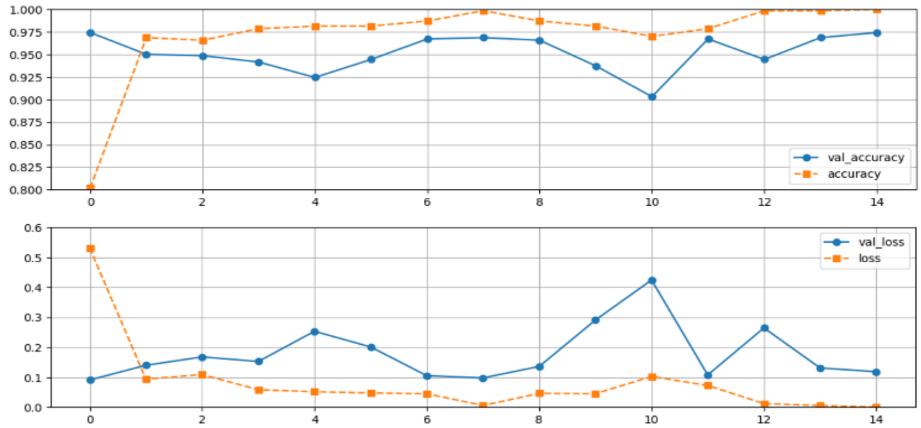
Tabel 1. Hasil Pengujian Model ViT

Label	Precision	Recall	F1-Score
0 (Batun Poh)	72%	86%	79%
1 (Kekandikan)	85%	73%	79%
2 (Kojongan)	88%	93%	90%
3 (Sesrengatan)	84%	82%	83%
3 (Beblatungan)	94%	71%	81%
Akurasi	82%		

Berdasarkan hasil pada Tabel 1 dapat ditunjukkan bahwa akurasi model secara keseluruhan hanya 82% dan data yang memiliki *f1-score* paling rendah adalah warangka jenis Batun Poh dan Kekandikan.

2) Model *ConvNeXt*

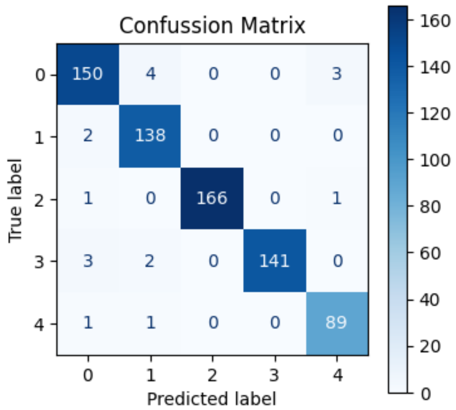
Setelah melakukan pelatihan pada 15 iterasi, didapatkan trend akurasi yang menunjukan peningkatan yang signifikan, sementara trend loss menunjukan penurunan. Hal ini mengindikasikan bahwa proses pembelajaran pada model *ConvNeXt* berlangsung dengan baik. Grafik unjuk kerja akurasi dan *loss* selama proses *training* pada model *ConvNeXt* dapat dilihat pada Gambar 12.



Gambar 12. Trend Akurasi (atas) dan *Loss* (bawah) Pada Model *ConvNeXt*

Grafik pada Gambar 12 menunjukan bahwa akurasi pelatihan mengalami peningkatan yang sangat cepat sejak awal, dan stabil di atas 97% mulai dari *epoch* ke-3. Peningkatan ini bahkan lebih cepat dari model ViT. Sementara *loss* pelatihan menurun tajam selama tiga epoch pertama dan kemudian terus menurun perlahan hingga mendekati nol. *Loss* pada data validasi menunjukkan sedikit fluktuasi, namun sebagian besar tetap berada dalam kisaran rendah (dibawah 0,2). Hal ini mengindikasikan proses pelatihan yang efisien dan kestabilan model dalam mengenali struktur data. Hal ini menunjukan bahwa *ConvNeXt* tidak hanya mempelajari data pelatihan dengan baik, tetapi juga mempertahankan stabilitas prediksi terhadap data validasi.

Setelah proses *training* selesai, kemudian berikutnya akan dilakukan proses pengujian pada model *ConvNeXt* yang sudah dilatih. Tujuannya untuk mengetahui unjuk kerja akhir model pada data pengujian. Proses ini dilakukan dengan menggunakan *data testing* yang sudah disiapkan. *Confusion matrix* dari hasil pengujian model *ConvNeXt* dapat dilihat pada Gambar 13.



Gambar 13. *Confusion Matrix* Model *ConvNeXt*

Dari Gambar 13 dapat dilihat bahwa jumlah data yang berhasil diklasifikasikan dengan baik oleh *ConvNeXt* jauh lebih banyak daripada model ViT. Meskipun begitu, *error* masih terlihat pada beberapa data namun dengan jumlah yang lebih sedikit. Kemudian dari confusion matrix pada Gambar 13 dapat dihitung akurasi, *precision, recall dan f1-score* nya. Hasilnya dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pengujian Model *ConvNeXt*

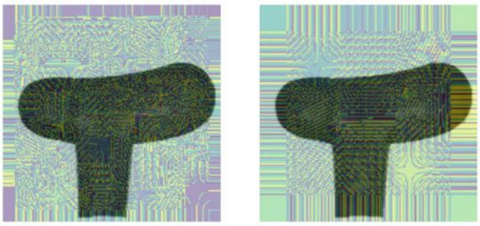
Label	Precision	Recall	F1-Score
0 (Batun Poh)	96%	96%	96%
1 (Kekandikan)	95%	99%	97%
2 (Kojongan)	100%	99%	99%
3 (Sesrengatan)	100%	97%	98%
4 (Beblatungan)	96%	98%	97%
Akurasi	97%		

Berdasarkan hasil pada Tabel 2 dapat ditunjukkan bahwa akurasi model secara keseluruhan mencapai 97%. Nilai ini jauh lebih tinggi dari akurasi yang dihasilkan oleh ViT.

3) Eksperimen

Eksperimen dilakukan untuk menunjukan apakah model dapat fokus untuk mendeteksi objek tertentu pada fitur yang bersifat diskriminatif. Semakin terfokus, maka semakin baik model dalam memahami konteks gambar. Metode yang digunakan adalah *Grad CAM*. Hasilnya dapat dilihat pada Gambar 14.

Visualisasi *Grad-CAM* pada Gambar 14 (kiri) yang menggunakan model ViT dan Gambar 14 (kanan) yang menggunakan model *ConvNeXt* menunjukkan perbedaan signifikan dalam pola aktivasi spasial terhadap objek warangka keris. Model ViT memperlihatkan distribusi aktivasi yang lebih menyebar, termasuk pada area latar belakang, yang mengindikasikan bahwa model ini belum sepenuhnya mampu memfokuskan perhatian hanya pada fitur diskriminatif dari objek utama. Hal ini dapat menyebabkan sensitivitas terhadap noise dan potensi kesalahan klasifikasi. Sebaliknya, model *ConvNeXt* menunjukkan pola aktivasi yang lebih terfokus pada bagian tengah objek, khususnya pada area yang merepresentasikan bentuk utama warangka keris. Aktivasi yang terlokalisasi dengan baik ini menandakan kemampuan *ConvNeXt* dalam mengidentifikasi fitur representatif secara lebih selektif dan efisien, sekaligus menunjukkan generalisasi spasial yang lebih baik.



Gambar 14. Hasil Eksperimen *Grad CAM* Pada Model ViT (kiri) dan model *ConvNeXt* (kanan)

4) Pembahasan

Berdasarkan hasil pelatihan dan pengujian, model *ConvNeXt* memiliki kinerja lebih baik daripada model *Vision Transformers* (ViT). Hal ini ditinjau dari unjuk kerja training, pengujian dengan *confusion matrix*, akurasi model, hingga hasil eksperimen dengan menggunakan Grad CAM. Perbandingan proses pelatihan antara model *Vision Transformer* (ViT) dan *ConvNeXt* menunjukkan dinamika konvergensi dan generalisasi yang berbeda secara signifikan. Pada grafik ViT (kiri), terlihat bahwa akurasi pelatihan meningkat tajam hingga mendekati 1, sementara akurasi validasi (val_accuracy) mengalami stagnasi di kisaran 0.75 setelah sekitar epoch ke-10. Sementara itu, nilai val_loss pada ViT tidak menunjukkan penurunan yang stabil, namun cenderung fluktuatif setelahnya. Sebaliknya, grafik ConvNeXt (kanan) memperlihatkan proses pelatihan yang lebih stabil dan efisien.

Berdasarkan *confusion matrix* hasil klasifikasi menggunakan model *Vision Transformer* (ViT), model menunjukkan unjuk kerja yang cukup baik pada beberapa kelas, meskipun masih terdapat kesalahan klasifikasi antar kelas. Kelas Kojongan (2) memiliki tingkat akurasi tertinggi dengan 173 citra yang berhasil diklasifikasikan dengan benar dari total keseluruhan instance-nya, serta sedikit kesalahan prediksi ke kelas lain. Demikian pula, kelas Batun Poh (0) juga menunjukkan unjuk kerja yang kuat dengan 144 prediksi yang benar, meskipun beberapa citra salah diklasifikasikan ke kelas Kekandikan (1) dan Sesrengatan (3). Untuk kelas Kekandikan (1), terdapat 94 prediksi yang benar, namun juga terlihat cukup banyak kekeliruan ke kelas Batun Poh dan Kojongan, yang mengindikasikan adanya kemiripan visual antar kelas tersebut. Sementara itu, kelas Sesrengatan (3) memiliki 108 prediksi yang benar, namun cukup sering tertukar dengan kelas Batun Poh dan Kekandikan, menunjukkan bahwa model mungkin mengalami kesulitan membedakan fitur visual antara kelas-kelas ini. Kelas Beblatungan (4) tampaknya merupakan kelas yang paling sulit dikenali oleh model, dengan hanya 63 prediksi yang benar dan sejumlah besar citra yang diklasifikasikan salah ke kelas Batun Poh, Kekandikan, dan Sesrengatan.

Hal ini menunjukkan bahwa representasi fitur untuk kelas Beblatungan kemungkinan besar kurang terserap dengan baik oleh model, atau bisa juga disebabkan oleh distribusi data yang tidak seimbang atau kemiripan visual yang tinggi dengan kelas lain. Secara keseluruhan, meskipun model menunjukkan kinerja yang baik pada kelas-kelas tertentu, diperlukan peningkatan, terutama dalam membedakan kelas-kelas yang memiliki karakteristik visual serupa. *Confusion matrix* dari model *ConvNeXt* menunjukkan dominasi prediksi yang benar pada setiap kelas, khususnya pada kelas Kojongan (label 2) dengan 166 prediksi benar, serta Sesrengatan (label 3) dan Batun Poh (label 0) dengan 141 dan 150 prediksi benar secara berturut-turut. Kesalahan klasifikasi antar kelas pada *ConvNeXt* juga sangat minim dan cenderung tidak menyebar, mengindikasikan kemampuan representasi spasial yang lebih tajam dalam membedakan ciri visual dari masing-masing warangka.

Hasil visualisasi Grad-CAM terhadap klasifikasi objek warangka keris Bali memperlihatkan perbedaan mendasar dalam distribusi atensi spasial antara model *Vision Transformer* (ViT) dan *ConvNeXt*. Grad-CAM pada model ViT menunjukkan pola aktivasi yang cenderung menyebar, termasuk pada area latar belakang gambar. Hal ini mengindikasikan bahwa ViT belum sepenuhnya mampu membedakan bagian-bagian penting dari objek utama secara spesifik. Aktivasi yang tidak terfokus dapat menyebabkan penurunan akurasi klasifikasi, terutama bila objek memiliki fitur visual yang mirip antar kelas. Sebaliknya, visualisasi Grad-CAM dari model *ConvNeXt* menunjukkan konsentrasi aktivasi yang lebih terarah dan terfokus pada area utama dari warangka keris, khususnya bagian tengah dan kontur dominan objek. Atensi yang terlokalisasi ini mencerminkan bahwa *ConvNeXt* berhasil mengenali fitur-fitur visual yang lebih relevan dan representatif terhadap label kelas. Fokus spasial yang baik ini memperkuat bukti bahwa ConvNeXt memiliki pemahaman fitur visual yang lebih dalam, dan lebih efisien dalam membedakan antar kelas berdasarkan informasi morfologis yang signifikan.

Penelitian ini memberikan perspektif baru terhadap efektivitas arsitektur deep learning dalam klasifikasi citra budaya. Berbeda dengan penelitian sebelumnya seperti oleh Sihananto *et al.* (2024) [17], Tran *et al.* (2025) [18], dan Pei *et al.* (2023) [19], yang menempatkan *Vision Transformer* (ViT) sebagai model pilihan utama, hasil penelitian ini menunjukkan bahwa *ConvNeXt* mampu memberikan kinerja yang lebih baik. *ConvNeXt* menunjukkan konvergensi pelatihan yang lebih stabil, kesalahan klasifikasi antar kelas yang lebih rendah, serta hasil visualisasi Grad-CAM yang lebih terfokus pada objek utama. Hal ini menunjukkan kemampuan representasi spasial yang lebih tajam dibandingkan ViT, khususnya dalam membedakan fitur visual halus antar kelas warangka keris Bali. Dengan demikian, penelitian ini memperkuat pemahaman bahwa arsitektur konvolusional modern seperti *ConvNeXt* tetap relevan dan bahkan mampu menghasilkan kinerja lebih baik dalam konteks klasifikasi citra budaya lokal yang kompleks.

4. Simpulan

Berdasarkan hasil pembahasan, dapat disimpulkan bahwa secara keseluruhan *ConvNeXt* menunjukkan keunggulan yang jelas dalam aspek kestabilan pelatihan, kecepatan konvergensi, dan kemampuan generalisasi dibandingkan dengan ViT. Hal ini dapat ditunjukkan dari akurasi model yang tinggi pada *ConvNeXt* mencapai 97%, dibandingkan dengan model ViT yang hanya menyentuh angka 82%. Dari hasil Grad CAM juga menunjukan bahwa *ConvNeXt* berhasil memberikan atensi yang lebih baik dalam mendeteksi bagian terpenting benda. Dengan demikian, dapat disimpulkan bahwa *ConvNeXt* tidak hanya unggul dalam unjuk kerja klasifikasi secara numerik, tetapi juga lebih efisien dalam mengalokasikan perhatian model ke bagian-bagian penting dari citra, yang merupakan indikator penting dalam evaluasi model berbasis interpretabilitas. Hal ini dapat dikaitkan dengan arsitektur *ConvNeXt* yang menggabungkan kekuatan konvolusi modern dengan efisiensi distribusi perhatian spasial, menjadikannya lebih adaptif terhadap data visual kompleks seperti warangka keris Bali.

Referensi

[1] Q. Xuanhao and Z. Min, "A Review of Attention Mechanisms in Computer Vision," in *Proceedings of the 2023 8th International Conference on Image, Vision and Computing (ICIVC)*, 2023, pp. 577–583, doi: 10.1109/ICIVC58118.2023.10270435.

[2] A. Vaswani *et al*., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998-6008, 2017.

[3] X. Yang, "An Overview of the Attention Mechanisms in Computer Vision," *J. Phys. Conf. Ser.*, vol. 1693, p. 12173, 2020, doi: 10.1088/1742-6596/1693/1/012173.

[4] H. Li and N. Chen, "The Use of Computer Vision Technology in the Inheritance of Intangible Cultural Heritage: The Case of Regional Cultural Characteristics," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, pp. 1–20, Jan. 2024, doi: 10.2478/amns-2024-1853.

[5] L. Silva, O. Bellon, and K. Boyer, "Computer Vision And Graphics For Heritage Preservation And Digital Archaeology," *Revista de Informática Teórica e Aplicada*, vol. 11, pp. 9–32, 2004, doi: 10.22456/2175-2745.5746.

[6] J. Mitrić, I. Radulović, T. Popović, Z. Šćekić, dan S. Tinaj, "AI and Computer Vision in Cultural Heritage Preservation," *Proc. 2024 28th International Conference on Information Technology (IT)*, Žabljak, Montenegro, 21–24 Jan. 2024, pp. 1–4, doi: 10.1109/IT61232.2024.10475738.

[7] B. Dana, "Identitas Seniman Tari Barong dan Keris terhadap Komodifikasi Tari Sakral di Batubulan," *Mudra: Jurnal Seni Budaya*, vol. 37, no. 3, pp. 265–270, 2022, doi: 10.31091/mudra.v37i3.1702. [8] B. C. Mintaraga, *Desain Keris Bali kontemporer: kajian makna simbolis \& filosofis*. 2019.

[9] A. Fahrurrozhi and H. Kurnia, "Memahami Kekayaan Budaya dan Tradisi Suku Bali di Pulau Dewata yang Menakjubkan," *Jurnal Ilmu Sosial dan Budaya Indonesia*, vol. 2, no. 1, pp. 39–50, 2024, doi: 10.61476/6635j851.

[10] I. M. Made Ardika Yasa, Ida Bagus Putu Arnyana, dan I. Wayan Suastra, "Keris sebagai representatif manusia dalam peradaban masyarakat Bali di Lombok," *Widya Sandhi*, vol. 14, no. 2, Nov. 2023, pp. 88–107, doi: 10.53977/ws.v14i2.1078.

[11] I. Sujana, "Legal Implications of Keris Marriage on the Inheritance Rights of Balinese Women: A Human Rights Perspective," *Sci. Law*, vol. 2025, pp. 43–47, 2025, doi: 10.55284/s2e1cs64.

[12] I. G. M. D. Hartawan and K. A. Wiratni, "Pande Besi Di Era Modern (Studi di Desa Sawan, Kecamatan Sawan, Kabupaten Buleleng)," *J. Daya Saing*, vol. 9, no. 3, pp. 665–674, 2023, doi: 10.35446/dayasaing.v9i3.1456.

[13] Desa Bengkala, "Statistik Berdasar Pekerjaan," 2025. https://bengkala-buleleng.desa.id/index.php/first/statistik/pekerjaan (accessed May 16, 2025).

[14] Desa Bukti, "Statistik Desa Bukti Berdasar Pekerjaan," 2025. https://bukti-buleleng.desa.id/index.php/first/statistik/pekerjaan (accessed May 16, 2025).

[15] Desa Sanggalangit, "Statistik Berdasar Pekerjaan," 2025. https://sanggalangit-buleleng.desa.id/index.php/first/statistik/pekerjaan (accessed May 10, 2025).

[16] N. J. W. Park, H. Regenbrecht, S. Duncan, S. Mills, R. W. Lindeman, N. Pantidi, dan H. Whaanga, "Mixed Reality Co-Design for Indigenous Culture Preservation & Continuation," *Proceedings of the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 149–157, Mar. 2022, doi: 10.1109/VR51125.2022.00033 .

[17] A. N. Sihananto, M. M. Al Haromainy, Z. A. Fauzi, R. A. Reza, G. C. H. Putra, and T. M. Christianty, "Wayang's Images Recognition using Vision Transformer," *Int. J. Data Sci. Eng. Anaylitics*, vol. 4, no. 2, pp. 15–27, 2024, doi: 10.33005/ijdasea.v4i2.24.

[18] M.-T. Tran, T.-P. Pham, T.-N. Nguyen, and T.-N. Do, "Classifying Intangible Cultural Heritage Images in the Mekong Delta," *SN Comput. Sci.*, vol. 6, no. 6, p. 584, 2025, doi: 10.1007/s42979-025-04117-8.

[19] H. Pei, C. Zhang, X. Zhang, X. Liu, and Y. Ma, "Recognizing Materials In Cultural Relic Images Using Computer Vision And Attention Mechanism," *Expert Syst. Appl.*, vol. 239, p. 122399, 2024, doi: https://doi.org/10.1016/j.eswa.2023.122399.

[20] T. Fan, H. Wang, and S. Deng, "Intangible Cultural Heritage Image Classification With Multimodal Attention And Hierarchical Fusion," *Expert Syst. Appl.*, vol. 231, p. 120555, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120555.

[21] L. Gao, Y. Wu, T. Yang, X. Zhang, Z. Zeng, C. K. D. Chan, and W. Chen, "Research on Image Classification and Retrieval Using Deep Learning with Attention Mechanism on

Diaspora Chinese Architectural Heritage in Jiangmen, China," *Buildings*, vol. 13, no. 2, art. no. 275, pp. 1–21, Jan. 2023, doi: 10.3390/buildings13020275.
[22] J. Ahmad, K. Muhammad, and S. Baik, "Data Augmentation-Assisted Deep Learning Of Hand-Drawn Partially Colored Sketches For Visual Search," *PLoS One*, vol. 12, p. e0183838, 2017, doi: 10.1371/journal.pone.0183838.
[23] J. Yao, L. Xing, and H. Wu, "A Microblog Content Credibility Evaluation Model Based On The Influence Of Sentiment Polarity," *Mobile Information Systems*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/8983534.
[24] H. Kato, K. Osuge, S. Haruta, and I. Sasase, "A Preprocessing by Using Multiple Steganography for Intentional Image Downsampling on CNN-Based Steganalysis," *IEEE Access*, vol. 8, pp. 195578–195593, 2020, doi: 10.1109/ACCESS.2020.3033814.
[25] K. Alrfou, A. Kordijazi, and T. Zhao, "Computer Vision Methods for the Microstructural Analysis of Materials: The State-of-the-art and Future Perspectives." 2022, doi: 10.48550/arXiv.2208.04149.
[26] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-based Image Captioning," *arXiv preprint* arXiv:2107.06912, pp. 1–27, Jul. 2021, doi: 10.48550/arXiv.2107.06912.
[27] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Prepr. arXiv2010.11929*, 2020.
[28] S. Kirstein, H. Wersing, H.-M. Gross, and E. Körner, "A Vector Quantization Approach For Life-Long Learning Of Categories" in *Proc. 18th Int. Conf. Artificial Neural Networks (ICANN)*," vol. 5506, pp. 805–812, 2008, doi: 10.1007/978-3-642-02490-0_98.
[29] V. Singla, S. Bawa, and J. Singh, "Enhancing Indian Sign Language Recognition Through Data Augmentation And Visual Transformer," *Neural Comput. Appl.*, vol. 36, pp. 1–14, 2024, doi: 10.1007/s00521-024-09845-1.
[30] T. Zhang, W. Xu, B. Luo, and G. Wang, "Depth-Wise Convolutions In Vision Transformers For Efficient Training On Small Datasets" *Neurocomputing*, vol. 617, p. 128998, 2025.
[31] R. Ibadulla, T. M. Chen, and C. C. Reyes-Aldasoro, "ConvShareViT: Enhancing Vision Transformers with Convolutional Attention Mechanisms for Free-Space Optical Accelerators," *arXiv Prepr. arXiv2504.11517*, 2025.
[32] B. Graham *et al.*, "Levit: A Vision Transformer In Convnet's Clothing For Faster Inference" in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12259–12269.
[33] A. Todi, N. Narula, M. Sharma, and U. Gupta, "ConvNext: A Contemporary Architecture for Convolutional Neural Networks for Image Classification," 2023, pp. 1–6, doi: 10.1109/CISCT57197.2023.10351320.
[34] Z. Xing, Y. Liu, Q. Wang, and J. Fu, "Fault Diagnosis Of Rotating Parts Integrating Transfer Learning And Convnext Model, " *Sci. Rep.*, vol. 15, no. 1, p. 190, 2025, doi: 10.1038/s41598-024-84783-5.
[35] Z. Li, T. Gu, B. Li, W. Xu, X. He, and X. Hui, "ConvNeXt-Based Fine-Grained Image Classification and Bilinear Attention Mechanism Model," *Appl. Sci.*, vol. 12, no. 18, 2022, doi: 10.3390/app12189016.
[36] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A Convnet For The 2020s," In *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*, 2022, pp. 11976–11986.
[37] D. Setiawan, A. S. Karnyoto, I. Intan, and B. Pardamean, "ConvNeXt Model for Breast Cancer Image Classification," in *Proc. 2024 6th International Conference on Radar, Informatics and Systems (ICORIS)*, Surakarta, Indonesia, Nov. 2024, pp. 1–5, doi: 10.1109/ICORIS63540.2024.10903832.
[38] S. Tufchi, A. Yadav, and T. Ahmed, "AMTCF: An Advanced Multimodal Transformer And Convnext Fusion For Contextualized Fake News Detection In Digital Landscape," *Lang. Resour. Eval.*, pp. 1–35, 2025, doi: 10.1007/s10579-025-09838-z.
[39] A. Ameneshewa, "ConvNeXt Based Hybrid Models with Multi-Modal Feature Fusion for ECG Classification," in *Artificial Intelligence and Human-Computer Interaction*, Mar. 2025, pp. 166–175, doi: 10.1007/978-981-97-3965-0_14.
[40] Y. Zhang, A. Xu, D. Lan, X. Zhang, J. Yin, and H. H. Goh, "Convnext-Based Anchor-Free Object Detection Model For Infrared Image Of Power Equipment," *Energy Reports*, vol. 9, pp. 1121–1132, Sep. 2023, doi: 10.1016/j.energyrep.2023.04.145.

[41] L. Ramos, E. Casas, C. Romero, F. Rivas, and M. E. Morocho-Cayamcela, “A Study Of Convnext Architectures For Enhanced Image Captioning,” *IEEE Access*, vol. 12, pp. 13711-13728., 2024, doi: 10.1109/ACCESS.2024.3356551.
[42] M. Zhao, X. Xu, X. Bao, X. Chen, and H. Yang, “An Automated Instance Segmentation Method For Crack Detection Integrated With Crackmover Data Augmentation,” *Sensors*, vol. 24, no. 2, p. 446, 2024, doi: 10.3390/s24020446.