Program Studi Teknik Informatika
Institut Teknologi Sumatera

Nama: **Ferdana Al Hakim (122140012)** Tugas Ke: **Laporan Perbandingan Vision Transformer**

Mata Kuliah: **Deep Learning** Tanggal: 21 November 2025

GitHub Repository:
http://github.com/luciferdana/VisionTransformer-Comparison

1 Pendahuluan

Vision Transformer (ViT) telah merevolusi bidang computer vision dengan mengadaptasi arsitektur Transformer yang awalnya dirancang untuk Natural Language Processing (NLP) ke dalam domain pemrosesan gambar [1]. Berbeda dengan Convolutional Neural Networks (CNN) tradisional yang mengandalkan operasi konvolusi lokal, Vision Transformer membagi gambar menjadi patch-patch kecil dan memprosesnya menggunakan mekanisme self-attention, memungkinkan model untuk menangkap dependensi global dalam gambar.

Penelitian ini bertujuan untuk membandingkan performa tiga arsitektur Vision Transformer yang populer: Vision Transformer (ViT) [1], Swin Transformer [2], dan Data-efficient Image Transformer (DeiT) [3] pada dataset CIFAR-100 [4]. Perbandingan dilakukan berdasarkan beberapa metrik evaluasi, termasuk akurasi, presisi, recall, F1-score, waktu inferensi, dan throughput.

Dataset CIFAR-100 terdiri dari 60.000 gambar berwarna berukuran 32x32 piksel yang terbagi dalam 100 kelas, dengan masing-masing kelas memiliki 600 gambar. Dataset ini merupakan tantangan yang menarik untuk menguji kemampuan model Vision Transformer dalam klasifikasi gambar multi-kelas dengan resolusi rendah.

2 Metodologi

2.1 Dataset dan Preprocessing

Eksperimen ini menggunakan dataset CIFAR-100 yang terdiri dari 50.000 gambar untuk pelatihan dan 10.000 gambar untuk pengujian. Setiap gambar memiliki dimensi 32x32 piksel dan termasuk dalam salah satu dari 100 kelas yang berbeda.

Untuk preprocessing data, dilakukan beberapa transformasi:

- **Data Training:** Resize ke 224x224, Random Crop dengan padding 4, Random Horizontal Flip, normalisasi dengan mean=[0.5071, 0.4867, 0.4408] dan std=[0.2675, 0.2565, 0.2761]
- **Data Testing:** Resize ke 224x224, normalisasi dengan mean dan std yang sama dengan data training

2.2 Arsitektur Model

Tiga arsitektur Vision Transformer yang dibandingkan dalam penelitian ini adalah:

2.2.1 Vision Transformer (ViT)

ViT adalah arsitektur Transformer murni untuk klasifikasi gambar yang diperkenalkan oleh Dosovitskiy et al. [1]. Model ini membagi gambar input menjadi patch berukuran 16x16 piksel, kemudian setiap patch diproyeksikan ke dalam embedding space dan diproses menggunakan Transformer encoder standar. Model yang digunakan adalah `vit_base_patch16_224` dengan total 85.88 juta parameter.

2.2.2 Swin Transformer

Swin Transformer adalah arsitektur hierarkis yang menggunakan shifted windows untuk menghitung self-attention [2]. Pendekatan ini mengurangi kompleksitas komputasi dari kuadratik menjadi linear terhadap ukuran gambar. Model yang digunakan adalah `swin_base_patch4_window7_224` dengan total 86.85 juta parameter.

2.2.3 Data-efficient Image Transformer (DeiT)

DeiT adalah varian dari ViT yang dirancang untuk lebih efisien dalam hal data training [3]. Model ini menggunakan teknik distilasi pengetahuan (knowledge distillation) untuk meningkatkan performa dengan data training yang lebih sedikit. Model yang digunakan adalah `deit_base_patch16_224` dengan total 85.88 juta parameter.

2.3 Konfigurasi Training

Semua model dilatih dengan konfigurasi yang sama untuk memastikan perbandingan yang adil:

- **Epochs:** 5
- **Batch Size:** 32
- **Learning Rate:** 1e-4
- **Optimizer:** AdamW dengan weight decay 0.01
- **Scheduler:** Cosine Annealing Learning Rate
- **Loss Function:** Cross Entropy Loss
- **Hardware:** Tesla P100-PCIE-16GB GPU

2.4 Metrik Evaluasi

Performa model dievaluasi menggunakan metrik-metrik berikut:

- **Accuracy:** Persentase prediksi yang benar dari total prediksi
- **Precision:** Rasio true positive terhadap total prediksi positif (weighted average)
- **Recall:** Rasio true positive terhadap total sampel positif aktual (weighted average)
- **F1-Score:** Harmonic mean dari precision dan recall (weighted average)
- **Inference Time:** Waktu rata-rata untuk memproses satu gambar (dalam milidetik)
- **Throughput:** Jumlah gambar yang dapat diproses per detik (frames per second)

3 Hasil Eksperimen

3.1 Performa Model

Tabel 1 menunjukkan ringkasan performa dari ketiga model Vision Transformer yang diuji pada dataset CIFAR-100.

Tabel 1: Ringkasan Hasil Eksperimen

Model	Parameters (M)	Size (MB)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (ms)
ViT	85.88	327.59	88.80	89.08	88.80	88.83	5.71
Swin Transformer	86.85	331.29	**92.05**	**92.17**	**92.05**	**92.06**	6.52
DeiT	85.88	327.59	89.74	89.92	89.74	89.75	**5.71**

Dari Tabel 1, dapat dilihat bahwa **Swin Transformer** mencapai akurasi tertinggi sebesar 92.05%, diikuti oleh DeiT dengan 89.74% dan ViT dengan 88.80%. Swin Transformer juga mengungguli model lainnya dalam metrik precision, recall, dan F1-score. Namun, dalam hal kecepatan inferensi, ViT dan DeiT lebih unggul dengan waktu inferensi 5.71 ms dibandingkan Swin Transformer yang memerlukan 6.52 ms.

3.2 Visualisasi Training History

Gambar 1 menunjukkan grafik training dan validation loss serta accuracy selama 5 epoch training untuk ketiga model.

Dari Gambar 1, dapat diamati bahwa semua model menunjukkan penurunan loss yang konsisten dan peningkatan akurasi selama training. Namun, terdapat gap yang signifikan antara training accuracy dan validation accuracy, mengindikasikan adanya overfitting. Swin Transformer menunjukkan validation accuracy yang paling stabil dan tertinggi di antara ketiga model.

3.3 Perbandingan Metrik Klasifikasi

Gambar 2 menunjukkan perbandingan visual dari metrik-metrik evaluasi (accuracy, precision, recall, F1-score) untuk ketiga model.

Swin Transformer secara konsisten mengungguli kedua model lainnya di semua metrik klasifikasi. Perbedaan performa antara Swin Transformer dan DeiT adalah sekitar 2.3%, sedangkan perbedaan dengan ViT adalah sekitar 3.3%.

3.4 Perbandingan Parameter dan Ukuran Model

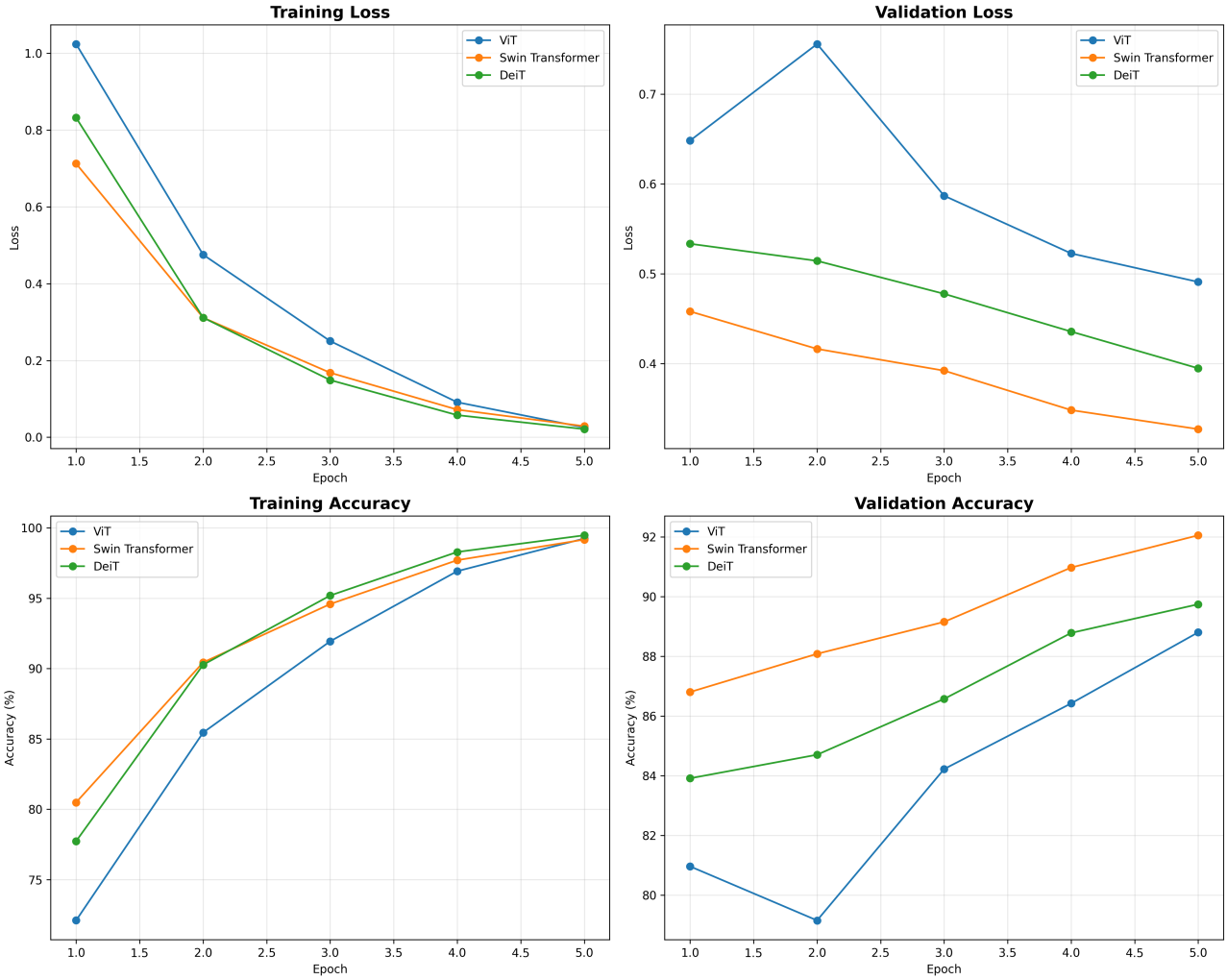
Gambar 3 menunjukkan perbandingan jumlah parameter dan ukuran model dalam megabytes.

Swin Transformer memiliki jumlah parameter sedikit lebih banyak (86.85M) dibandingkan ViT dan DeiT (85.88M). Perbedaan ini tidak signifikan (sekitar 1.1%), namun berkontribusi pada peningkatan performa yang cukup substansial.

3.5 Perbandingan Waktu Inferensi

Gambar 4 menunjukkan perbandingan waktu inferensi rata-rata dan throughput untuk ketiga model.

ViT dan DeiT memiliki waktu inferensi yang identik (5.71 ms) dengan throughput sekitar 175 fps. Swin Transformer sedikit lebih lambat dengan waktu inferensi 6.52 ms dan throughput 153.40 fps. Perbedaan kecepatan ini (sekitar 14%) dapat diterima mengingat peningkatan akurasi yang signifikan.



Gambar 1: Training History untuk ViT, Swin Transformer, dan DeiT

3.6 Confusion Matrix

Gambar 5, 6, dan 7 menunjukkan confusion matrix untuk masing-masing model pada dataset testing.

Dari confusion matrix dapat diamati bahwa Swin Transformer memiliki distribusi prediksi yang lebih terkonsentrasi pada diagonal (prediksi benar), mengindikasikan performa klasifikasi yang lebih baik dibandingkan ViT dan DeiT.

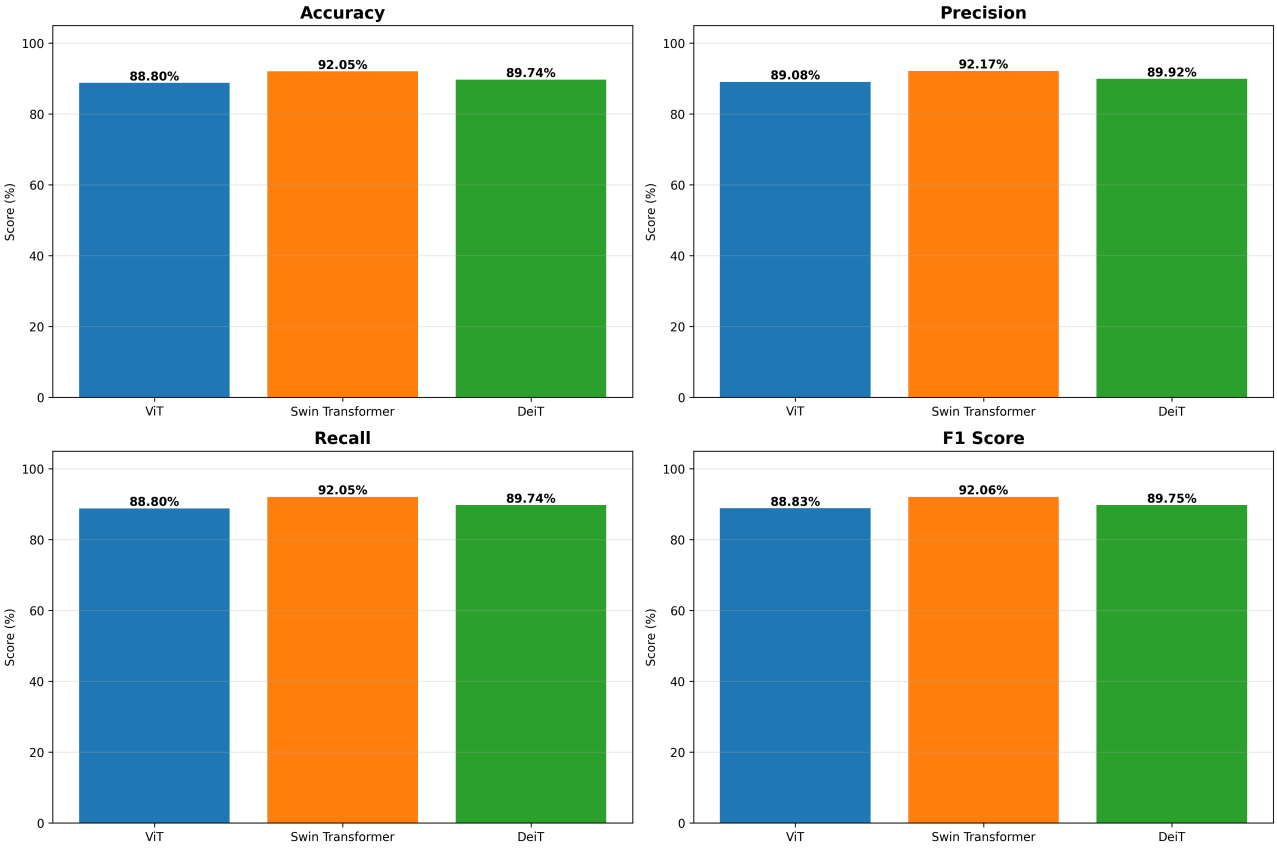
4 Pembahasan

4.1 Analisis Performa Model

Hasil eksperimen menunjukkan bahwa Swin Transformer mencapai performa terbaik dengan akurasi 92.05% pada dataset CIFAR-100. Keunggulan Swin Transformer dapat dijelaskan oleh beberapa faktor:

1. **Hierarchical Architecture:** Swin Transformer menggunakan arsitektur hierarkis yang memungkinkan model untuk menangkap fitur pada berbagai skala, mirip dengan CNN tradisional namun dengan keunggulan self-attention mechanism.

2. **Shifted Window Attention:** Mekanisme shifted window mengurangi kompleksitas komputasi



Gambar 2: Perbandingan Metrik Klasifikasi

sambil tetap mempertahankan kemampuan untuk menangkap dependensi global. Hal ini memungkinkan model untuk lebih efisien dalam memproses informasi spasial.

3. **Inductive Bias:** Meskipun Transformer umumnya memiliki inductive bias yang lebih sedikit dibandingkan CNN, arsitektur hierarkis Swin Transformer memberikan struktur yang lebih cocok untuk data visual.

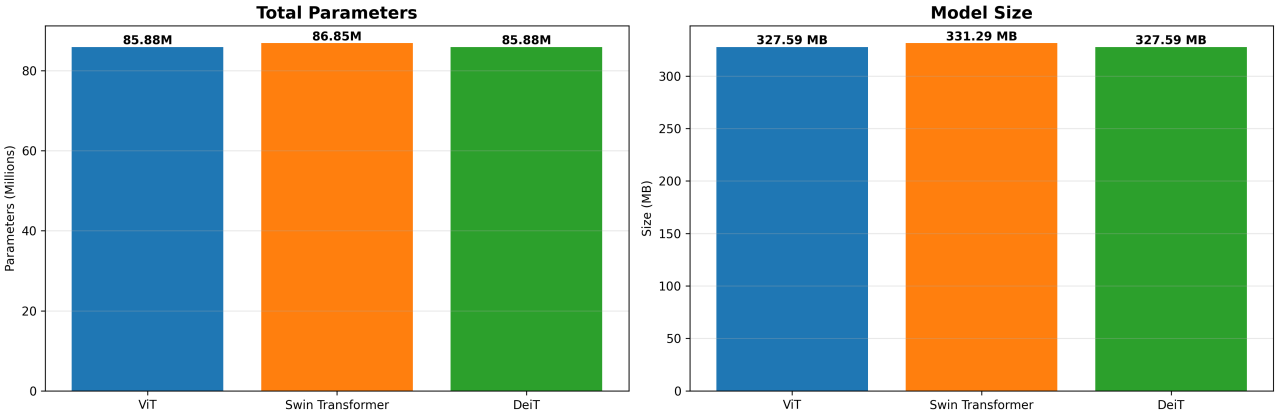
ViT, meskipun merupakan arsitektur yang lebih sederhana dan lebih cepat, menunjukkan performa yang sedikit lebih rendah (88.80%). Hal ini kemungkinan disebabkan oleh kurangnya inductive bias spasial yang dimiliki oleh arsitektur hierarkis seperti Swin Transformer.

DeiT menunjukkan performa yang berada di tengah-tengah (89.74%), yang konsisten dengan desainnya untuk efisiensi data. Meskipun DeiT dirancang untuk bekerja dengan baik pada dataset yang lebih kecil, pada eksperimen ini dengan CIFAR-100 yang memiliki 50.000 sampel training, keunggulannya tidak terlalu signifikan dibandingkan ViT.

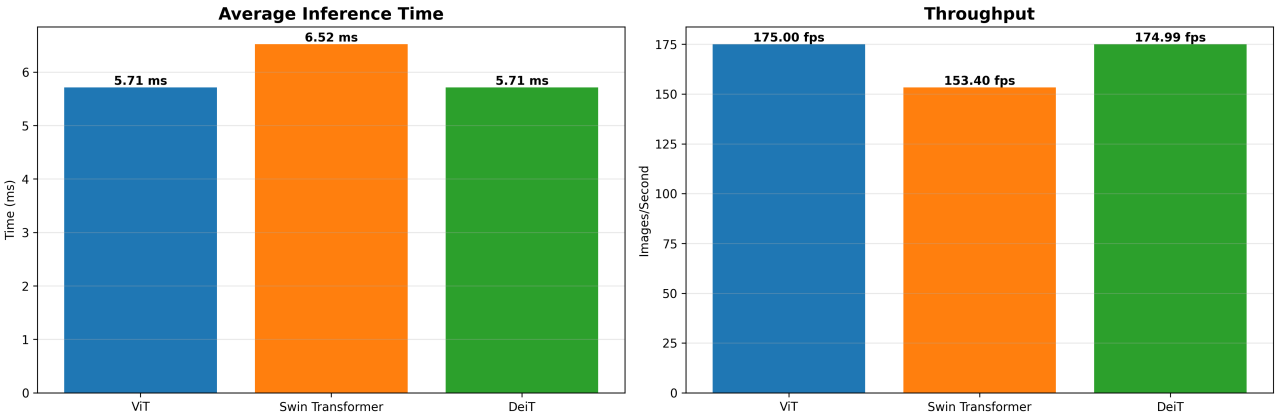
4.2 Trade-off Akurasi vs Kecepatan

Terdapat trade-off yang jelas antara akurasi dan kecepatan inferensi:

- Swin Transformer menawarkan akurasi tertinggi (92.05%) namun dengan kecepatan inferensi yang sedikit lebih lambat (6.52 ms)
- ViT dan DeiT menawarkan kecepatan inferensi yang lebih cepat (5.71 ms) namun dengan akurasi yang sedikit lebih rendah



Gambar 3: Perbandingan Parameter dan Ukuran Model



Gambar 4: Perbandingan Waktu Inferensi dan Throughput

Perbedaan waktu inferensi sekitar 14% antara Swin Transformer dan ViT/DeiT dapat diterima untuk aplikasi yang mengutamakan akurasi, seperti medical imaging atau quality control. Sebaliknya, untuk aplikasi real-time seperti autonomous driving atau video surveillance, ViT atau DeiT mungkin lebih cocok.

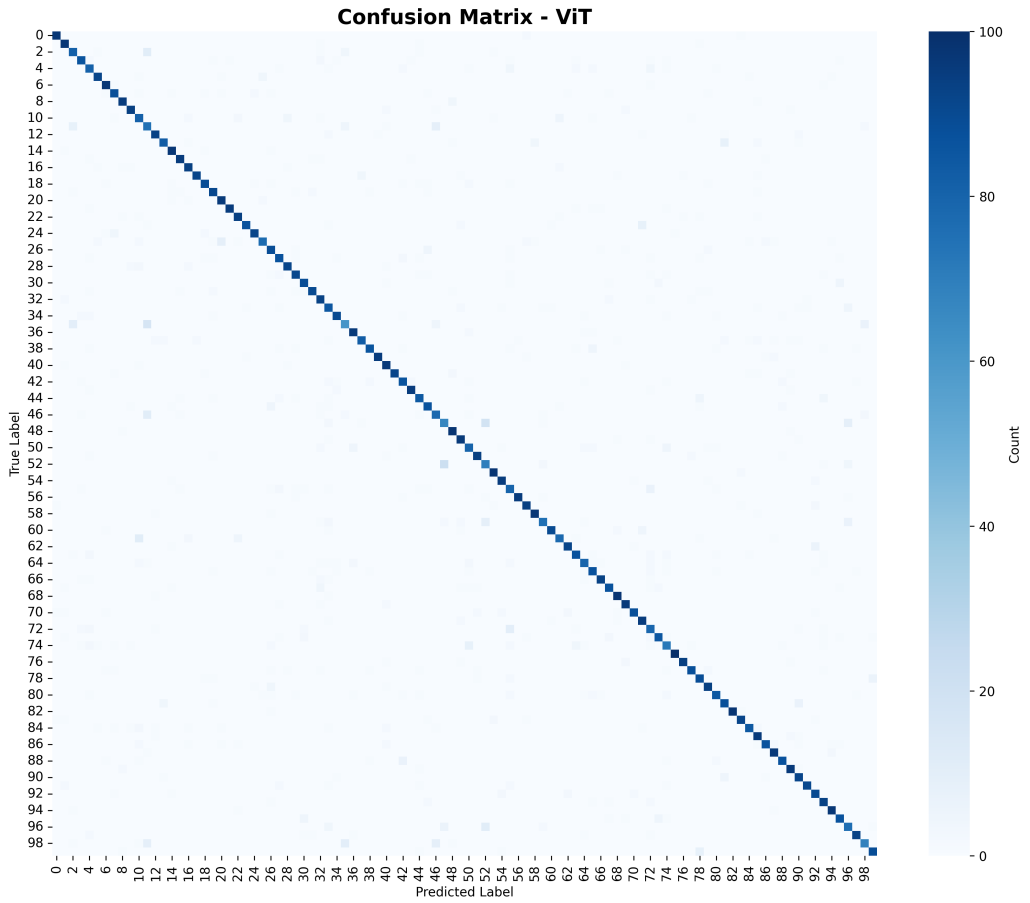
4.3 Overfitting

Semua model menunjukkan tanda-tanda overfitting yang signifikan, dengan training accuracy mencapai 99% sementara validation accuracy berada di kisaran 88-92%. Beberapa strategi yang dapat digunakan untuk mengurangi overfitting:

- Augmentasi data yang lebih agresif
- Regularisasi yang lebih kuat (weight decay, dropout)
- Early stopping berdasarkan validation loss
- Training dengan epoch yang lebih sedikit
- Menggunakan teknik seperti mixup atau cutmix

4.4 Efisiensi Parameter

Ketiga model memiliki jumlah parameter yang relatif serupa (85-87 juta parameter), namun menunjukkan performa yang berbeda. Hal ini mengindikasikan bahwa arsitektur dan bagaimana parameter



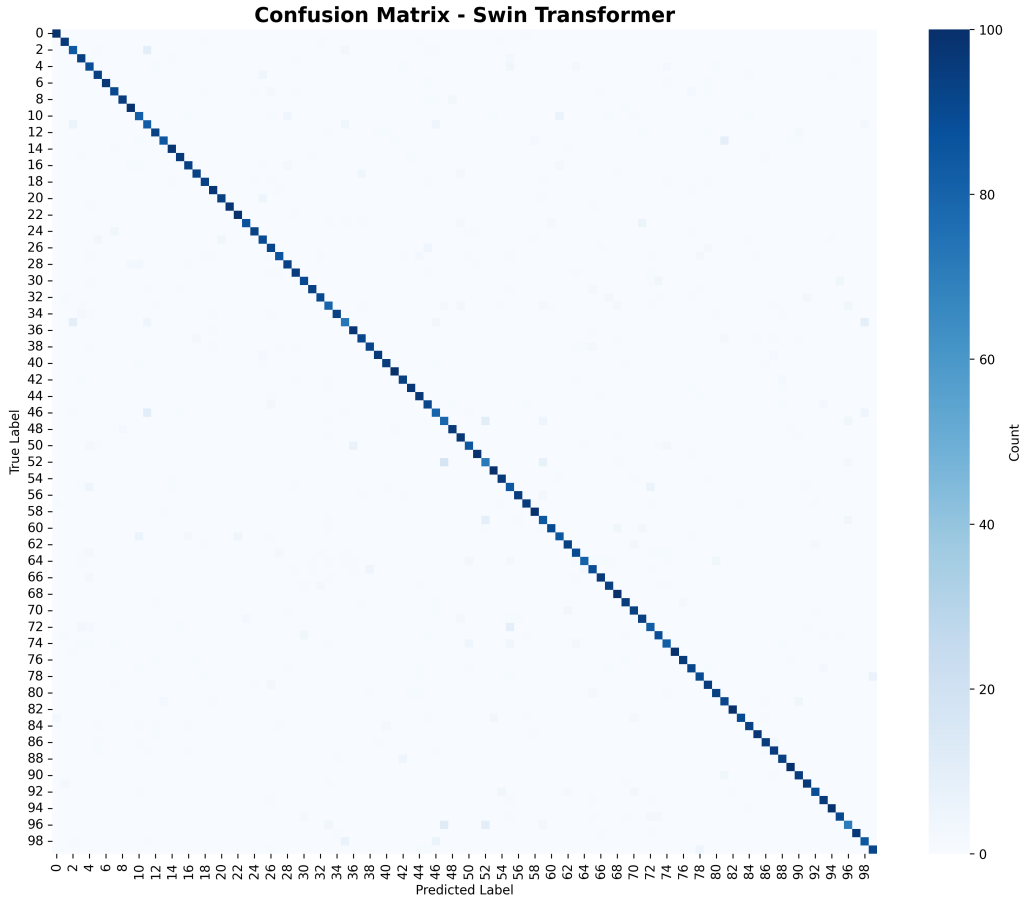
Gambar 5: Confusion Matrix - Vision Transformer (ViT)

tersebut diorganisir lebih penting daripada jumlah parameter semata. Swin Transformer, dengan hanya 1.1% lebih banyak parameter dibandingkan ViT, mampu mencapai peningkatan akurasi sebesar 3.3%.

5 Kesimpulan

Penelitian ini membandingkan tiga arsitektur Vision Transformer (ViT, Swin Transformer, dan DeiT) pada dataset CIFAR-100. Berdasarkan hasil eksperimen, dapat disimpulkan bahwa:

1. **Swin Transformer** mencapai performa terbaik dengan akurasi 92.05%, precision 92.17%, recall 92.05%, dan F1-score 92.06%, mengungguli ViT dan DeiT secara signifikan.
2. **ViT dan DeiT** menawarkan kecepatan inferensi yang lebih tinggi (5.71 ms atau 175 fps) dibandingkan Swin Transformer (6.52 ms atau 153.40 fps), dengan perbedaan sekitar 14%.
3. Terdapat **trade-off antara akurasi dan kecepatan**: Swin Transformer lebih akurat namun sedikit lebih lambat, sementara ViT dan DeiT lebih cepat namun dengan akurasi yang sedikit lebih rendah.
4. Semua model menunjukkan **overfitting yang signifikan**, mengindikasikan perlunya strategi regularisasi yang lebih baik atau augmentasi data yang lebih agresif.
5. **Arsitektur hierarkis** dari Swin Transformer terbukti lebih efektif untuk klasifikasi gambar dibandingkan arsitektur Transformer murni seperti ViT dan DeiT, terutama pada dataset dengan resolusi rendah seperti CIFAR-100.



Gambar 6: Confusion Matrix - Swin Transformer

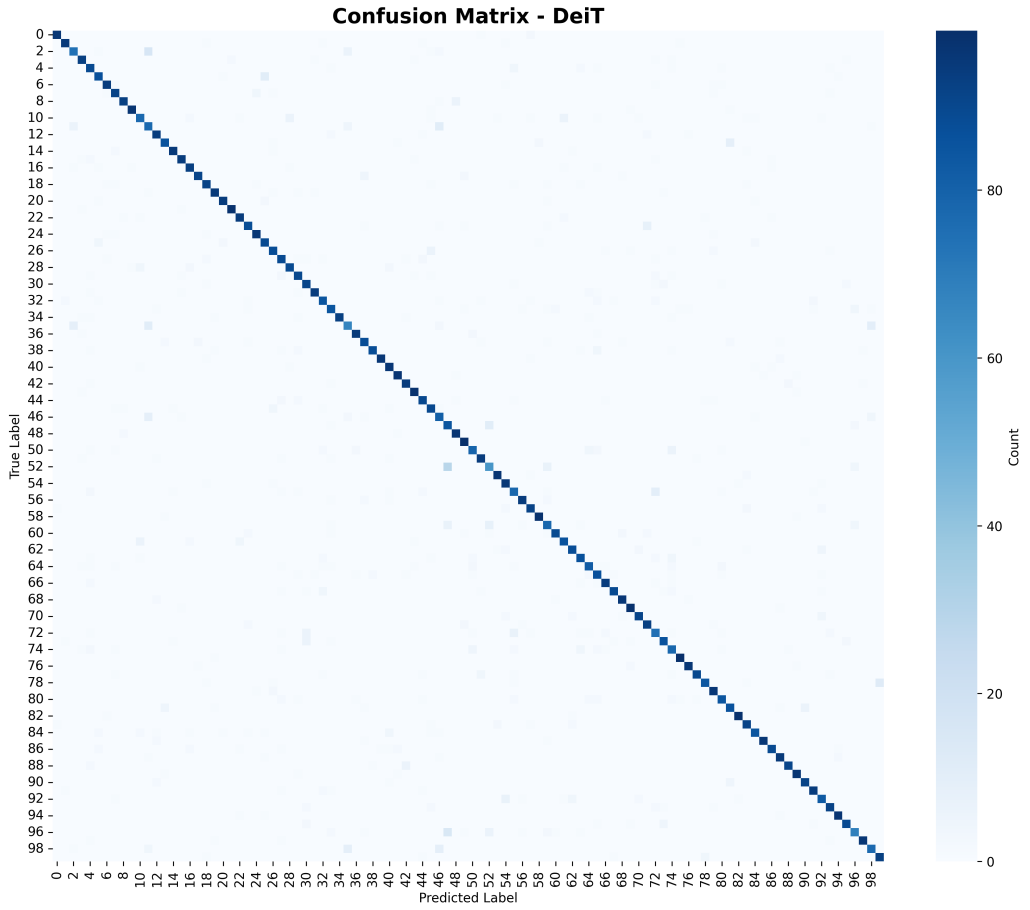
References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[3] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[4] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Technical report*, 2009.



Gambar 7: Confusion Matrix - Data-efficient Image Transformer (DeiT)