ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO TỔNG KẾT ĐỒ ÁN

ĐỀ TÀI: PHÂN LOẠI CẢNH QUAN

Nhóm sinh viên thực hiện:

Cáp Kim Hải Anh - 23520036

Hồ Ngọc Luật - 23520900

Nguyễn Phạm Phương Nam - 23520978

Lớp: CS231.P21.KHTN

Giảng viên: TS. Mai Tiến Dũng

Ngày 23 tháng 06 năm 2025

Mục lục

1 Thông tin các thành viên và mức độ hoàn thành:

- Hồ Ngọc Luật - 23520900
  - Triển khai mô hình EfficientViT.
  - Slide thuyết trình, nội dung báo cáo.
  - Mức độ hoàn thành: 100%.
- Cáp Kim Hải Anh - 23520036
  - Triển khai mô hình EfficientNetV2.
  - Slide thuyết trình, nội dung báo cáo.
  - Mức độ hoàn thành: 100%.
- Nguyễn Phạm Phương Nam - 23520978
  - Triển khai mô hình Swin Tiny.
  - Slide thuyết trình, nội dung báo cáo.
  - Mức độ hoàn thành: 100%.

2 Tổng quan và Lý do lựa chọn bài toán:

2.1 Tổng quan:

- Với sự phát triển vượt bậc của công nghệ hình ảnh và trí tuệ nhân tạo đã thúc đẩy mạnh mẽ nghiên cứu về phân loại ảnh, trong đó, phân loại cảnh quan (scene classification) là một nhiệm vụ cốt lõi và đầy thách thức. Đây là bài toán nhằm gán nhãn loại cảnh phù hợp cho một bức ảnh đầu vào dựa trên các đặc điểm tổng thể thay vì chỉ nhận diện các vật thể riêng lẻ bên trong. Nhiệm vụ này đòi hỏi mô hình phải hiểu ngữ cảnh rộng, mối quan hệ không gian và đặc trưng cấu trúc toàn cục để phân biệt các loại cảnh như đô thị, nông thôn, tự nhiên (rừng, biển, núi) hoặc khu vực chức năng (sân bay, cảng,...).
- Ngoài ra, với sự phát triển mạnh mẽ của học sâu, đặc biệt là các kiến trúc mạng nơ-ron tích chập (CNN) và gần đây là các mô hình dựa trên Transformer, khả năng tự động phân loại cảnh quan đã đạt được những bước tiến vượt bậc.

2.2 Lý do lựa chọn bài toán:

- Trong bối cảnh dữ liệu ảnh từ vệ tinh và thiết bị bay không người lái ngày càng phổ biến, việc tự động phân loại ảnh cảnh quan từ trên cao (scene classification) giúp máy tính có thể nhận biết và hiểu được nội dung tổng quát của một khung cảnh dựa trên đặc điểm

hình ảnh như kiến trúc, địa hình. Nó không chỉ là một bài toán học thuật quan trọng mà còn mang lại giá trị ứng dụng to lớn trong:

  - Giám sát đô thị và quy hoạch đất đai.
  - Quản lý tài nguyên thiên nhiên và môi trường.
  - Ứng phó thiên tai và cứu hộ.
  - Phát hiện và theo dõi thay đổi địa lý.
  - Xây dựng hệ thống bản đồ tự động thông minh.

- Ngoài ra, các cảnh quan thường rất đa dạng về hình thái, có thể là sự khác biệt giữa các lớp rất tinh tế (ví dụ: khu dân cư đông đúc và khu dân cư vừa), hoặc nhiều lớp có hình thái thị giác tương đồng (ví dụ: sân vận động và trường học), đồng thời là điều kiện chụp ảnh không đồng nhất. Vì vậy, đòi hỏi các phương pháp tốt để phân loại.

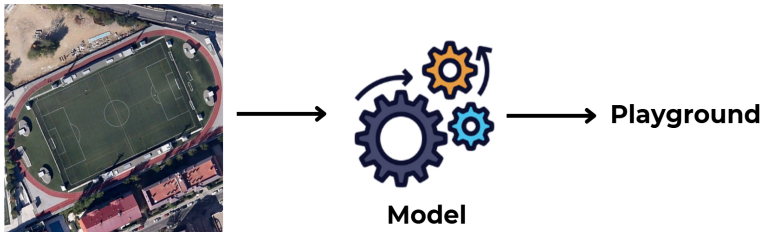
2.3 Mục tiêu:

- Tìm hiểu và phân tích được đặc trưng của dữ liệu.
- Nghiên cứu và triển khai các mô hình học sâu hiện đại để giải quyết bài toán phân loại ảnh cảnh quan.
- Huấn luyện và đánh giá hiệu suất của các mô hình đã chọn trên bộ dữ liệu thực nghiệm. Đánh giá hiệu quả của mô hình thông qua các chỉ số như độ chính xác, thời gian, ...
- So sánh kết quả thực nghiệm để xác định tính hiệu quả của mỗi phương pháp trong bài toán phân loại cảnh quan.
- Rút ra đánh giá, nhận xét tổng kết toàn bộ quá trình thực hiện đồ án.

3 Phát biểu bài toán:

- **Tên bài toán:** Phân loại cảnh quan (Scene Classification).
- **Input**:
  - Tập dữ liệu cảnh quan đã được gán nhãn, bao gồm:
    * Tập các nhãn.
    * Mỗi phần tử của tập dữ liệu gồm một số là ảnh cảnh quan và nhãn được gán tương ứng cho ảnh đó.
  - Ảnh số x là ảnh cảnh quan cần dự đoán.
- **Output:** Giá trị nhãn dự đoán của ảnh x, thuộc tập các nhãn.

- **Điều kiện:** Ảnh được chụp từ vệ tinh, flycam hoặc drone (ảnh chụp trên không). Ảnh chỉ có 1 cảnh quan.



4 Phương pháp Xử lý dữ liệu:

4.1 Biến đổi dữ liệu:

- **Mục tiêu:** Chuẩn bị dữ liệu cho quá trình huấn luyện mô hình. Giúp mô hình học được các đặc trưng mạnh mẽ hơn, giảm thiểu hiện tượng quá khớp (overfitting) và cải thiện khả năng tổng quát hóa của mô hình. Đặc biệt đối với ảnh cảnh quan, sự đa dạng về góc chụp, ánh sáng và điều kiện môi trường là rất lớn.
- **Các kỹ thuật được áp dụng:**
  - **Resize:** Chuyển đổi kích thước ảnh về chuẩn đầu vào của mô hình (thường là 224 x 224), đảm bảo tính nhất quán.
  - **Random Rotation:** Xoay ảnh một góc ngẫu nhiên trong khoảng 30 độ, tăng tính linh hoạt trước sự thay đổi góc nhìn.
  - **Horizontal Flip:** Lật ảnh theo chiều ngang với xác suất 50%, giúp mô hình nhận diện đối tượng bất kể hướng trái - phải.
  - **Color Jitter:** Thay đổi độ sáng, độ tương phản, độ bão hòa và sắc độ của ảnh, giúp mô hình ổn định hơn với điều kiện ánh sáng đa dạng trong môi trường thực tế.
  - **Random Crop:** Cắt ngẫu nhiên một vùng trong ảnh, sau đó resize về 224×224. Điều này giúp mô hình tập trung vào nhiều khu vực khác nhau trong ảnh.
  - **Normalize:** Chuẩn hóa giá trị pixel của ảnh về phân phối chuẩn với các giá trị trung bình và độ lệch chuẩn giống như của ImageNet. Điều này giúp tăng tốc độ hội tụ của quá trình huấn luyện và cải thiện hiệu suất.

Tất cả những bước trên đều được thực hiện thông qua thư viện torchvision.transforms trong PyTorch

4.2 Chia tập dữ liệu:

- Để đảm bảo huấn luyện mô hình một cách công bằng và tránh hiện tượng lệch phân phối, toàn bộ tập dữ liệu được chia thành ba tập: huấn luyện (train), kiểm định (validation) và kiểm tra (test) với tỷ lệ tương ứng là 60% - 20% - 20%.
- Mỗi ảnh được gán nhãn sẵn dựa trên cấu trúc thư mục của bộ dữ liệu gốc, được nạp qua ImageFolder. Ảnh của mỗi lớp được gom lại, sau đó chia đều theo đúng tỷ lệ. Một bước điều chỉnh được thực hiện để đảm bảo tổng số ảnh sau chia không vượt hoặc thiếu so với ban đầu do làm tròn số lượng ảnh.
- Sau khi chia, ảnh được sao chép sang thư mục tương ứng: train, val và test, mỗi lớp nằm trong thư mục con riêng để phù hợp với chuẩn đầu vào của ImageFolder khi load lại về sau.
- Cách làm này đảm bảo sự đồng đều của phân phối lớp trong cả ba tập, giúp mô hình học tổng quát tốt hơn và phản ánh đúng hiệu năng khi đánh giá.

5 Phương pháp giải quyết:

5.1 Mô hình EfficientNetV2:

5.1.1 Tổng quan mô hình:

- EfficientNetV2 là một trong những kiến trúc mạng tích chập (CNN) được phát triển bởi Google, kế thừa và cải tiến từ EfficientNet. Mục tiêu chính là đạt được hiệu suất cao nhất với chi phí tính toán thấp nhất, giải quyết các hạn chế của EfficientNet. Mô hình này được thiết kế dựa trên kỹ thuật Compound Scaling.
- Mô hình EfficientNetV2 kết hợp hai loại khối:
  - **Fused-MBConv** (khối mới): thay thế khối MBConv truyền thống bằng cách tích hợp convolution chuẩn và depthwise trong một bước, dùng ở các tầng đầu để tăng tốc tính toán và suy luận, đồng thời vẫn giữ được hiệu quả tham số.
  - **MBConv:** khối convolution mở rộng, chuẩn hóa và thu gọn giúp trích xuất đặc trưng sâu hơn, sử dụng ở các tầng cuối.
- Cấu trúc chia thành các stage từ 0 đến 6, mỗi stage học đặc trưng từ mức thấp đến mức cao. Cuối cùng, mô hình sử dụng global pooling và fully connected layer để phân loại.

5.1.2 Huấn luyện mô hình:

- **Khởi tạo mô hình:** mô hình EfficientNetV2 được khởi tạo từ thư viện timm với biến thể **'efficientnetv2_s'**, sử dụng mô hình có dùng trọng số pretrained và mô hình không dùng trọng số pretrained để đánh giá.
- **Phần cuối của mỗi mô hình được tuỳ chỉnh đầu ra:** model.classifier được thay thế bằng một head mới gồm:
  - Lớp linear đầu: Giảm chiều không gian xuống 512.
  - BatchNorm: Ổn định hoá đầu ra (đảm bảo mỗi tầng có dữ liệu ổn định để học, tránh đầu ra các lớp liên tục thay đổi (phóng đại, làm lệch).
  - ReLU: Tăng tính phi tuyến.
  - Dropout: Tắt ngẫu nhiên 50% neurons, không để các neron phối hợp quá chặt chẽ, ép mỗi neuron phải học độc lập, giúp giảm overfitting.
  - Lớp linear cuối: Tuỳ chỉnh đầu ra bằng số lớp bộ dữ liệu sử dụng (num_classes)
- **Cấu hình mô hình huấn luyện:**
  - **Hàm mất mát:** Cross-entropy với Label Smoothing (0.1): Giảm độ tin tưởng tuyệt đối vào nhãn đúng, giúp mô hình tránh quá khớp và tăng khả năng tổng quát.
  - **Tối ưu hóa:** sử dụng thuật toán AdamW là phiên bản cải tiến của Adam với tách biệt L2 regularization, giúp hội tụ tốt hơn, đặc biệt trong mạng sâu. Với learning rate khởi tạo 1e-4 và trọng số giảm dần (weight decay) là 1e-5.
  - **Bộ scheduler OneCycleLR** lên kế hoạch điều chỉnh learning rate trong suốt quá trình huấn luyện với max_lr = 1e-3, sử dụng chiến lược 'cosine annealing', chia nhỏ thành 200 epoch với pct_start = 0.1: Khởi đầu thấp, tăng dần lên max, rồi giảm theo cosine. Giúp mô hình hội tụ nhanh và ổn định, giảm thời gian huấn luyện.
  - **Automatic Mixed Precision (AMP):** Giảm tiêu tốn bộ nhớ GPU mà không ảnh hưởng đáng kể đến độ chính xác.
  - **Gradient Clipping:** Giới hạn độ lớn của gradient ở mức 1.0 để tránh hiện tượng gradient exploding.
  - **Early Stopping:** Được thiết lập với patience = 10, tự động dừng khi độ chính xác trên tập validation không còn cải thiện, giúp tiết kiệm thời gian và tránh overfitting.
  - **Checkpoint:** Giúp khôi phục từ chỗ dừng giữa chừng (resume) hoặc phục vụ đánh giá, inference sau này.
- **Thông số huấn luyện:**
  - Số epoch: 200.
  - Batch size 32.

5.2 Mô hình Swin Transformer:

5.2.1 Tổng quan mô hình:

- Swin Transformer là kiến trúc Vision Transformer cải tiến, cho phép mô hình xử lý ảnh ở độ phân giải lớn với chi phí tính toán hợp lý. Các đặc điểm chính:
  - Hierarchical structure: tạo biểu diễn phân cấp tương tự CNN (giống ResNet), dễ tích hợp với các bài toán thị giác.
  - Window-based MSA (W-MSA): chia ảnh thành các vùng nhỏ (patch/window) không chồng lấn và tính self-attention cục bộ.
  - Shifted Window MSA (SW-MSA): cơ chế dịch cửa sổ giữa các lớp liên tiếp giúp kết nối các vùng khác nhau, tạo khả năng mô hình hóa toàn cục mà vẫn tiết kiệm tính toán.

5.2.2 Huấn luyện mô hình:

- **Khởi tạo mô hình:** Sử dụng mô hình Swin-Tiny có dùng pretrained và mô hình huấn luyện từ đầu.
- Phần head cuối gồm: Dropout 0.2 và Linear(num_features, num_classes).
- **Cấu hình mô hình huấn luyện:** Sử dụng loss và optimizer tương tự EfficientNetV2.
- **Thông số huấn luyện:**
  - Số epoch: 100.
  - Batch size 32.

5.3 Mô hình EfficientViT:

5.3.1 Tổng quan mô hình:

- EfficientViT là kiến trúc mới nhằm giải quyết ba vấn đề chính trong Vision Transformer:
  - Tối ưu bộ nhớ (Memory Efficiency): giảm số lớp MHSA (Multi-head Self Attention), tăng lớp FFN theo bố cục Sandwich.
  - Giảm dư thừa tính toán (Computation Redundancy): chia nhỏ input và truyền qua các đầu attention khác nhau để tăng tính đa dạng.
  - Tối ưu sử dụng tham số (Parameter Usage): sử dụng structured pruning để loại bỏ phần không quan trọng và phân bổ tài nguyên hiệu quả.

5.3.2 Huấn luyện mô hình:

- **Khởi tạo mô hình:** Sử dụng mô hình EfficientViT-M5 có dùng pretrained và mô hình không dùng pretrained.
- **Cấu hình mô hình huấn luyện:** tương tự như EfficientNetV2, sử dụng AdamW và scheduler OneCycleLR.
- **Thông số huấn luyện:**
  - Số epoch: 100 epoch.
  - Batch size 32.

6 Thực nghiệm và kết quả:

6.1 Tổng quan dataset sử dụng:

- **Tên bộ dữ liệu:** AID (Aerial Image Dataset) - là một bộ dữ liệu lớn và phổ biến trong lĩnh vực phân tích ảnh vệ tinh và viễn thám, được sử dụng rộng rãi để đánh giá hiệu suất của các thuật toán phân loại cảnh quan đô thị và tự nhiên. Bộ dữ liệu AID được phát triển bởi Đại học Vũ Hán, Trung Quốc. Hiện có trên kaggle tại đây.
- **Số mẫu:** 10 000 ảnh, với khoảng 200-400 ảnh cho mỗi lớp.
- **Bộ dữ liệu đa dạng với 30 lớp cảnh vật:** Các lớp này đại diện cho sự đa dạng của các loại địa hình và cấu trúc phổ biến trên bề mặt Trái Đất, từ các khu vực tự nhiên đến các khu vực nhân tạo (như sân bay, bãi biển, khu dân cư, đồng ruộng, sông, hồ,...).
- **Độ phân giải ảnh:** 600 x 600 pixel, định dạng RGB.
- **Nguồn ảnh:** được thu thập từ Google Earth, các vệ tinh thương mại và các nguồn công khai khác. Các ảnh được chụp từ nhiều quốc gia, nhiều mùa, nhiều thời điểm và độ cao khác nhau. Ảnh có độ phân giải cao và góc chụp thay đổi linh hoạt, tạo ra độ khó đáng kể cho bài toán phân loại. Điều này giúp tăng khả năng nghiên cứu chuyên sâu và thử nghiệm các mô hình học sâu hiện đại.
- **Một số đặc điểm nổi bật khác và lý do chọn dataset:**
  - Mỗi lớp chứa ảnh từ nhiều khu vực khác nhau, tạo nên sự đa dạng về điều kiện ánh sáng, môi trường và cấu trúc địa hình. Ảnh có độ phân giải cao, phù hợp cho việc huấn luyện các mô hình học sâu.
  - Độ chồng chéo về mặt trực quan giữa một số lớp (ví dụ: "industrial" và "commercial") làm tăng độ khó của bài toán, đòi hỏi mô hình phải học được đặc trưng ngữ nghĩa sâu sắc hơn.

6.2 Thực nghiệm:

- Các mô hình đều được huấn luyện trên nền tảng Kaggle Notebook GPU.
- Xây dựng lớp Dataset tuỳ chỉnh và tải bộ dữ liệu.
- Chia bộ dữ liệu thành 3 tập train (60%), validation (20%) và test (20%).
- Thực hiện biến đổi dữ liệu và khởi tạo DataLoader cho từng tập dữ liệu để nạp dữ liệu theo lô (batch) một cách hiệu quả.
- Khởi tạo và thiết lập cấu hình huấn luyện cho mỗi mô hình, sau đó thực hiện huấn luyện.
- Đánh giá mô hình dựa trên các chỉ số đánh giá. Các độ đo đánh giá:
  - **Accuracy (độ chính xác):** Tỉ lệ ảnh được phân loại đúng.
  - **Training Time:** Thời gian huấn luyện mô hình.
  - **Inference Time (CPU/GPU):** Thời gian dự đoán trung bình trên 1 ảnh.
  - **Model size:** Số lượng tham số/memory footprint của mô hình.

6.3 Kết quả thực nghiệm:

- Kết quả sau quá trình thực nghiệm của ba mô hình (sử dụng mô hình pretrained và có biến đổi dữ liệu) được thể hiện trong bảng dưới đây:

Tiêu chí	Accuracy	Time training	Inference CPU	Inference GPU	Model size
EfficientNetV2-S	95.20%	29m43s	94.95ms	4.12ms	20M
Swin Transformer	92.85%	1h43m	116.97ms	2.55ms	27M
EfficientViT-M5	94.74%	1h13m	18.46ms	1.67ms	12M

Bảng 1: Bảng so sánh kết quả trên ba phương pháp giải quyết

- Ngoài việc so sánh giữa các mô hình, nhóm cũng tiến hành đánh giá ảnh hưởng của hai yếu tố quan trọng đến độ chính xác của mô hình: trọng số pretrained và biến đổi dữ liệu đầu vào. Kết quả được trình bày ở bảng dưới đây:

Accuracy	Không biến đổi		Có biến đổi	
	Train từ đầu	Pretrained	Train từ đầu	Pretrained
EfficientNetV2-S	45.75%	93.80%	80.80%	95.20%
Swin Transformer	73.10%	94.60%	75.95%	92.85%
EfficientViT-M5	46.23%	92.06%	79.27%	94.74%

Bảng 2: Bảng so sánh độ chính xác giữa 3 mô hình trong toàn bộ 4 lần thử nghiệm

Ghi chú: Tất cả đều được thực hiện trên máy ảo của Kaggle.

6.4 Nhận xét kết quả:

6.4.1 So sánh tổng thể ba mô hình:

Từ kết quả thực nghiệm ở bảng 1:

- **Độ chính xác (Accuracy):**
  - Mô hình EfficientNetV2-S đạt độ chính xác cao nhất (95.20%), cho thấy hiệu quả của kiến trúc kết hợp giữa MBConv và Fused-MBConv khi áp dụng vào ảnh chụp từ trên cao.
  - EfficientViT-M5 cũng đạt độ chính xác khá cao (94.74%), nhỉnh hơn Swin Transformer (92.85%). Điều này cho thấy kiến trúc kết hợp attention và CNN của EfficientViT có khả năng học tốt với kích thước nhỏ gọn.
- **Thời gian huấn luyện:**
  - EfficientNetV2-S có thời gian huấn luyện ngắn nhất (29 phút 43 giây), cho thấy sự tối ưu cao về tốc độ huấn luyện.
  - Ngược lại, Swin Transformer mất nhiều thời gian nhất để hội tụ (hơn 1 giờ 40 phút), điều này có thể do đặc trưng attention yêu cầu nhiều tài nguyên tính toán.
- **Thời gian suy luận (Inference):**
  - Trên CPU, EfficientViT-M5 là nhanh nhất (18.46 ms), trong khi Swin Transformer chậm nhất (116.97 ms).
  - Trên GPU, khoảng cách giữa các mô hình được rút ngắn, nhưng EfficientViT-M5 vẫn vượt trội nhất (1.67 ms), phù hợp với ứng dụng thời gian thực.

6.4.2 Ảnh hưởng của biến đổi dữ liệu và pretrained:

Từ kết quả thực nghiệm ở bảng 2:

- **Trọng số pretrained** là yếu tố có ảnh hưởng rất rõ rệt đến kết quả: Cả ba mô hình đều có độ chính xác tăng mạnh khi sử dụng trọng số đã được huấn luyện trước (pretrained), đặc biệt là EfficientNetV2-S và EfficientViT-M5, với mức tăng hơn 45%. Điều này cho thấy pretrained giúp mô hình tận dụng được đặc trưng học từ các tập lớn như ImageNet, từ đó hội tụ nhanh hơn và đạt độ chính xác cao hơn trên tập dữ liệu AID.
- **Biến đổi dữ liệu (augmentation)** cũng có ảnh hưởng tích cực: Khi không dùng pretrained, việc áp dụng biến đổi dữ liệu giúp tăng đáng kể độ chính xác của các mô hình. Tuy nhiên, khi mô hình đã được pretrained, augmentation chỉ cải thiện nhẹ hoặc không đáng kể. Điều này có thể lý giải do đặc trưng học từ ImageNet đã đủ tổng quát và không còn lợi ích rõ ràng từ augmentation đơn giản.
- Toàn bộ kết quả cho thấy **pretrained weights và biến đổi dữ liệu** là yếu tố then chốt ảnh hưởng lớn đến hiệu quả mô hình học sâu, đặc biệt trong các bài toán phân loại ảnh.

7 Đánh giá, kết luận và mở rộng:

7.1 Đánh giá - Nhận xét:

- **Ưu điểm:**
  - EfficientNetV2-S là mô hình cân bằng tốt nhất giữa độ chính xác và tốc độ huấn luyện, trong khi EfficientViT-M5 phù hợp hơn cho môi trường giới hạn tài nguyên hoặc yêu cầu suy luận thời gian thực.
  - Việc lựa chọn và triển khai ba kiến trúc mô hình tiên tiến và khác biệt (EfficientNetV2 dựa trên CNN, Swin Transformer dựa trên Self-Attention, và EfficientViT kết hợp cả hai) cho phép một cái nhìn toàn diện về hiệu suất và đặc tính của các phương pháp học sâu hiện đại trong bài toán phân loại ảnh cảnh quan. Điều này giúp đánh giá được ưu điểm của từng loại kiến trúc trên bộ dữ liệu thực nghiệm AID.
  - Kết hợp nhiều kỹ thuật như OneCycleLR, AMP, EarlyStopping, ...
  - Khả năng mở rộng và tái dùng cao.
- **Hạn chế:**
  - Một số lớp dễ bị nhầm lẫn do tính chất thị giác tương đồng.
  - Chi phí huấn luyện cao với Transformer.
  - Chưa có sự kết hợp (ensemble) giữa các mô hình để tận dụng điểm mạnh riêng.

– Khả năng tổng quát hóa trên dữ liệu không nhìn thấy: Mặc dù đã áp dụng các kỹ thuật biến đổi dữ liệu, mô hình vẫn có thể gặp khó khăn khi gặp phải các biến thể cảnh quan mới, điều kiện ánh sáng khác biệt đáng kể.

7.2 Kết luận:

- Phân loại ảnh cảnh quan là một bài toán tiềm năng và khả thi với học sâu.
- Đồ án thực hiện việc nghiên cứu, triển khai và đánh giá các mô hình học sâu tiên tiến (EfficientNetV2, Swin Transformer, và EfficientViT) cho bài toán phân loại ảnh cảnh quan. Bằng cách sử dụng bộ dữ liệu thực nghiệm AID, nhóm đã tiến hành một cách có hệ thống quá trình tiền xử lý dữ liệu, huấn luyện mô hình, và đánh giá hiệu suất. Quá trình này không chỉ làm nổi bật khả năng mạnh mẽ của các kiến trúc này trong việc hiểu và phân loại các cảnh quan phức tạp từ ảnh vệ tinh mà còn cung cấp một cái nhìn định lượng về ưu điểm và nhược điểm của từng loại mô hình trong bối cảnh cụ thể của đề tài.
- Đồ án cũng đã đóng góp vào việc củng cố sự hiểu biết về các phương pháp học sâu mới nhất trong thị giác máy tính, từ các mạng tích chập hiệu quả đến các kiến trúc dựa trên Transformer.

7.3 Mở rộng:

- Áp dụng kỹ thuật Ensemble thay vì chỉ sử dụng một mô hình duy nhất để tận dụng thế mạnh của từng mô hình, từ đó cải thiện độ chính xác tổng thể và độ bền vững của hệ thống.
- Tối ưu tốc độ dự đoán và hiệu quả tính toán để triển khai thực tế hoặc có thể triển khai trên thiết bị có tài nguyên hạn chế.
- Thực hiện phân vùng ảnh (image segmentation) để cải thiện độ chính xác từng vùng.
- Kiểm tra và đánh giá hiệu suất của các mô hình đã huấn luyện trên các bộ dữ liệu ảnh cảnh quan khác (ngoài AID) để đánh giá khả năng tổng quát hóa của chúng.

8 Tài liệu tham khảo và Phụ lục:

Tài liệu

[1] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Devices," *arXiv preprint arXiv:1704.04861*, 2017. [Trực tuyến]. url: https://arxiv.org/abs/1704.04861.

[2] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017. [Trực tuyến]. url: https://arxiv.org/abs/1706.03762.

[3] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021. [Trực tuyến]. url: https://arxiv.org/abs/2010.11929.

[4] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proceedings of the International Conference on Machine Learning*, 2021. [Trực tuyến]. url: https://arxiv.org/abs/2104.00298.

[5] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. [Trực tuyến]. url: https://arxiv.org/abs/2103.14030.

[6] H. Han *et al.*, "EfficientViT: Multi-Scale Linear Attention for High-Resolution Dense Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. [Trực tuyến]. url: https://arxiv.org/abs/2305.07027.

[7] G.-S. Xia *et al.*, "AID: A Benchmark Dataset for Advanced Information Dissemination Remote Sensing Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 565-573, 2017. [Trực tuyến]. url: https://ieeexplore.ieee.org/document/7810793.

Phụ lục:

- **Source code cho kết quả ở Bảng 1:**
  - EfficientNetV2: Code.
  - Swin Transformer: Code.
  - EfficientViT: Code.
- **Source code cho kết quả ở Bảng 2:**
  - EfficientNetV2: Code.
  - Swin Transformer: Code.
  - EfficientViT: Code.
- **Slide thuyết trình:** Slide.