**Ứng dụng ConvNeXt và Vision Transformer trong phân loại hình ảnh do AI tạo và hình ảnh do con người tạo ra**

**Đồng Anh Quân\* ,Nguyễn Thị Quỳnh Anh\*, Thái Văn Sáng\*,**

**Vũ Thị Như Quỳnh\***

\*Đại học Thủy Lợi

***Tóm tắt:*** *Trước sự bùng nổ của các công cụ tạo sinh hình ảnh bằng trí tuệ nhân tạo như: DALL·E, Midjourney và Stable Diffusion, việc phân biệt giữa hình ảnh do AI tạo và hình ảnh do con người tạo trở thành một thách thức quan trọng trong lĩnh vực kiểm chứng nội dung số. Trong nghiên cứu này, chúng tôi sử dụng một tập dữ liệu đa dạng có sẵn gồm hai loại hình ảnh và áp dụng hai mô hình học sâu hiện đại là: Vision Transformer và ConvNext để thực hiện phân loại. Các mô hình được huấn luyện và đánh giá trên cùng một tập dữ liệu, sử dụng các kỹ thuật tiền xử lý và tăng cường dữ liệu. Kết quả cho thấy cả hai mô hình đều đạt độ chính xác cao, trong đó ConvNext nhỉnh hơn về hiệu suất (95%), cho thấy tiềm năng ứng dụng thực tế trong phát hiện hình ảnh do AI tạo. Nghiên cứu góp phần định hướng các giải pháp kiểm duyệt và xác thực hình ảnh trong thời đại AI hiện nay.*

## **GIỚI THIỆU**

Sự phát triển nhanh chóng của các mô hình trí tuệ nhân tạo tạo sinh (generative AI) trong lĩnh vực thị giác máy tính đã mang đến những thay đổi sâu sắc trong việc tạo ra nội dung hình ảnh. Các mô hình như: DALL-E 2 [1], Midjourney, và Stable Diffusion [2] có khả năng sinh ảnh từ văn bản với độ chân thực cao, thường khó phân biệt bằng mắt thường với các ảnh do con người chụp hoặc vẽ. Khả năng tạo ảnh chất lượng cao một cách dễ dàng đã mở ra nhiều ứng dụng sáng tạo trong nghệ thuật, truyền thông và thiết kế, nhưng đồng thời cũng đặt ra những rủi ro lớn liên quan đến thông tin sai lệch, giả mạo hình ảnh và thao túng dư luận [3].

Trong bối cảnh này, việc phát triển các phương pháp phân biệt hình ảnh do AI tạo với hình ảnh do con người tạo ra trở thành một yêu cầu cấp thiết. Các nghiên cứu gần đây đã chỉ ra rằng hình ảnh do AI tạo thường mang một số đặc điểm riêng biệt như: hiện tượng artefacts, bất thường trong kết cấu hoặc bố cục không hợp lý mà mô hình học sâu có thể khai thác [4]. Tuy nhiên, khi chất lượng hình ảnh từ các mô hình như Stable Diffusion 2.1 hay Midjourney ngày càng tiệm cận ảnh thật, việc phân biệt trở nên khó khăn hơn, đòi hỏi các kiến trúc mô hình mạnh mẽ và có khả năng trích xuất đặc trưng sâu.

Trong nghiên cứu này, chúng tôi xây dựng một hệ thống phân loại hình ảnh dựa trên hai kiến trúc học sâu hiện đại: Vision Transformer (ViT) [5] và ConvNeXt [6]. Hai mô hình đại diện cho hai hướng tiếp cận khác nhau trong học sâu, một bên dựa trên cơ chế attention và một bên dựa trên cơ chế tích chập. Chúng tôi sử dụng một tập dữ liệu có sẵn bao gồm các hình ảnh do AI tạo và hình ảnh do con người tạo, đồng thời huấn luyện và đánh giá các mô hình trên tập dữ liệu này.

Kết quả cho thấy cả hai mô hình đều đạt độ chính xác cao nhưng ConvNeXt cho hiệu năng nhỉnh hơn (95%). Nghiên cứu này không chỉ cung cấp một hướng tiếp cận hiệu quả cho bài toán nhận diện hình ảnh theo nguồn gốc mà còn góp phần vào nỗ lực xây dựng các công cụ kiểm chứng nội dung số, hướng tới môi trường truyền thông số minh bạch và đáng tin cậy.

## **II. CÁC NGHIÊN CỨU LIÊN QUAN**

#### **1. Nhận diện hình ảnh tạo sinh dựa trên đặc trưng hình học và nội dung**

Một số nghiên cứu ban đầu tập trung vào việc phát hiện các bất thường hình ảnh do mô hình tạo sinh gây ra, chẳng hạn như hiện tượng "artefacts", kết cấu không nhất quán hoặc chi tiết phi tự nhiên. Wang et al. [4] đã đề xuất một phương pháp phân tích miền tần số để phát hiện ảnh deepfake và ảnh do AI tạo, cho thấy rằng nhiều mô hình sinh ảnh để lại dấu vết trong phổ tần số. Tương tự, Zhang et al. [7] sử dụng các đặc trưng kết cấu (texture features) để phân biệt ảnh thực với ảnh sinh, tận dụng việc AI thường không tái tạo chính xác các chi tiết phức tạp như: tóc, mắt, hay bóng đổ.

#### **2. Ứng dụng mô hình học sâu để phát hiện ảnh tạo sinh**

Gần đây, các mô hình học sâu – đặc biệt là các kiến trúc CNN và Transformer – đã được áp dụng mạnh mẽ cho nhiệm vụ phát hiện ảnh do AI tạo. Trong [8], tác giả sử dụng ResNet50 được huấn luyện trên tập ảnh từ các mô hình GAN khác nhau để phát hiện ảnh giả, cho thấy mô hình có thể học được các đặc trưng sinh học không thể quan sát bằng mắt thường. Vision Transformer (ViT) [5] được xem là bước tiến quan trọng nhờ khả năng mô hình hóa quan hệ toàn cục giữa các vùng ảnh. ConvNeXt [6] là một kiến trúc CNN cải tiến, được thiết kế để đạt hiệu suất tương đương Transformer nhưng với lợi thế của mạng tích chập truyền thống. Cả hai mô hình này đều là đại diện tiêu biểu cho các hướng tiếp cận hiện đại trong bài toán phân loại ảnh.

#### **3. Đóng góp của nghiên cứu**

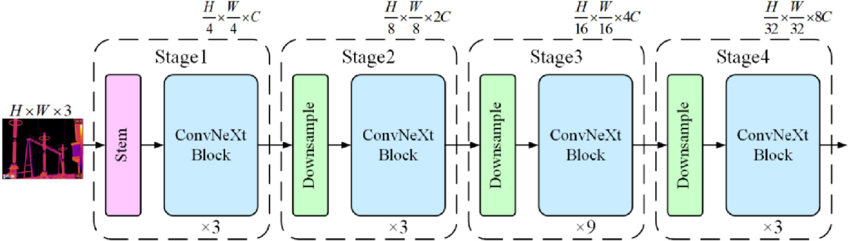
Mặc dù đã có nhiều phương pháp được đề xuất, vẫn còn thiếu các nghiên cứu đánh giá trực tiếp và so sánh hiệu quả giữa các kiến trúc hiện đại như ViT và ConvNeXt trong bài toán phân loại hình ảnh do AI tạo. Nghiên cứu này nhằm đánh giá hiệu quả hai kiến trúc học sâu hiện đại trên cùng một bài toán.

## **III. PHƯƠNG PHÁP NGHIÊN CỨU**

Trong nghiên cứu này, hai kiến trúc mạng sâu hiện đại là ConvNeXt và Vision Transformer (ViT) được triển khai và so sánh độc lập. Mỗi mô hình có cấu trúc đặc trưng riêng, phản ánh thế mạnh và nguyên lý hoạt động khác nhau. Việc sử dụng song song này nhằm đánh giá khả năng học đặc trưng, tổng quát hóa và hiệu quả phân loại trong bài toán phân biệt ảnh thật và ảnh AI.

#### ***1. Mô hình ConvNeXt***

ConvNext là một kiến trúc CNN hiện đại được thiết kế để cạnh tranh với các mô hình Transformer trong thị giác máy tính. ConvNeXt được xây dựng bằng cách hiện đại hóa ResNet, bao gồm các cải tiến như sử dụng các kernel tích chập lớn hơn, áp dụng chuẩn hóa Layer Normalization và sử dụng các kỹ thuật như stochastic depth và activation GELU. Kết quả là ConvNext đạt được hiệu suất cao trên các tác vụ như phân loại ảnh ImageNet, phát hiện đối tượng COCO và phân đoạn ngữ nghĩa ADE20K[9]. Mô hình đề xuất sử dụng ConvNeXt với các đặc điểm kiến trúc chính như sau:



*Hình 3.1 Tổng quan về kiến trúc ConvNext [9]*

*Stem:* Sử dụng khối tích chập lớn (4×4 hoặc 7×7) và stride lớn để giảm kích thước ảnh đầu vào.

*ConvNeXt Blocks:* Gồm nhiều khối depth-wise convolution kết hợp LayerNorm và skip connection.

*Activation:* Dùng GELU thay vì ReLU để học đặc trưng phi tuyến.

*Normalization:* Áp dụng LayerNorm thay cho BatchNorm.

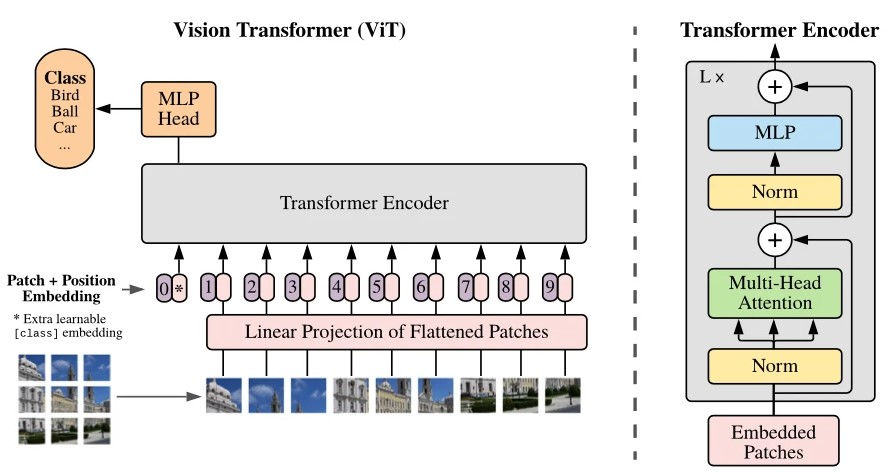
*Downsampling:* Thực hiện qua convolution stride 2 để giảm độ phân giải và tăng chiều sâu đặc trưng.

*Classifier Head:* Đầu ra qua Global Average Pooling và MLP để phân loại nhị phân (thật/giả).

Kiến trúc này tận dụng khả năng học đặc trưng không gian cục bộ mạnh mẽ, đồng thời cải tiến hiệu quả từ kinh nghiệm huấn luyện các mô hình lớn như Transformer.

#### ***2. Mô hình Vision Transformer (ViT)***

Vision Transformer (ViT) áp dụng mô hình Transformer từ xử lý ngôn ngữ tự nhiên vào thị giác máy tính. ViT chia hình ảnh thành các patch cố định, biến chúng thành vector và xử lý như một chuỗi đầu vào cho Transformer. Khi được huấn luyện trên các tập dữ liệu lớn như JFT-300M hoặc ImageNet-21k, ViT đạt hiệu suất vượt trội so với các mô hình CNN truyền thống như ResNet trên các benchmark như ImageNet, CIFAR-100 và VTAB [10]. Kiến trúc ViT đề xuất trong bài gồm các thành phần chính:



*Hình 3.2 Tổng quan về kiến trúc ViT [10]*

*Patch Embedding:* Chia ảnh thành patch (ví dụ 16×16), mỗi patch qua linear embedding thành vector.  
*Positional Encoding:* Thêm thông tin vị trí vào patch embedding.  
*Transformer Encoder:* Gồm nhiều lớp encoder với:

* Multi-Head Self-Attention (học quan hệ giữa các patch)
* Feed Forward Network (2 tầng MLP, GELU)
* Residual Connection và LayerNorm để ổn định và tăng hội tụ.
* CLS Token: Thêm token đặc biệt cho nhiệm vụ phân loại.
* MLP Head: Đưa vector CLS qua MLP cuối cùng để dự đoán nhãn.

ViT thể hiện khả năng tổng quát hóa vượt trội, đặc biệt trong việc nhận diện các kiểu mẫu toàn cục và các đặc trưng phân bố phi tuyến tính – điều thường thấy trong ảnh do AI tạo.

## **IV. KẾT QUẢ THỰC NGHIỆM**

**A - Bộ dữ liệu huấn luyện**

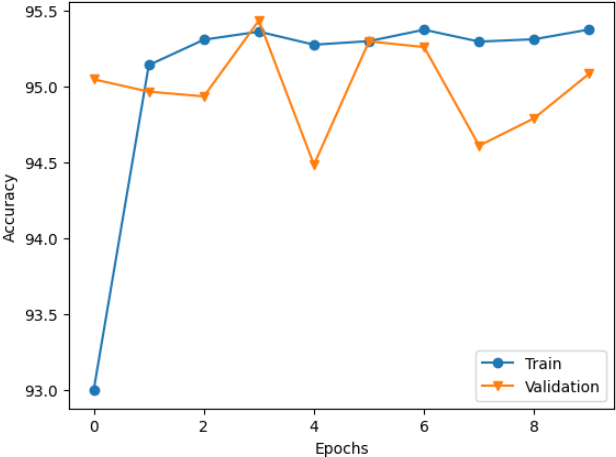
Nhóm sử dụng bộ dữ liệu "AI vs Human Generated Dataset" từ Kaggle gồm 80.000 ảnh huấn luyện và hơn 5.000 ảnh kiểm tra không nhãn. Ảnh được resize về 224×224, chuẩn hóa theo RGB và thống kê ImageNet, áp dụng các kỹ thuật tăng cường dữ liệu như: cắt, lật và điều chỉnh màu sắc.

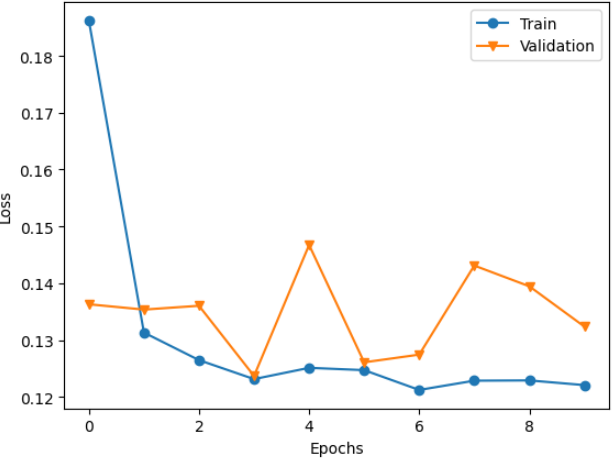


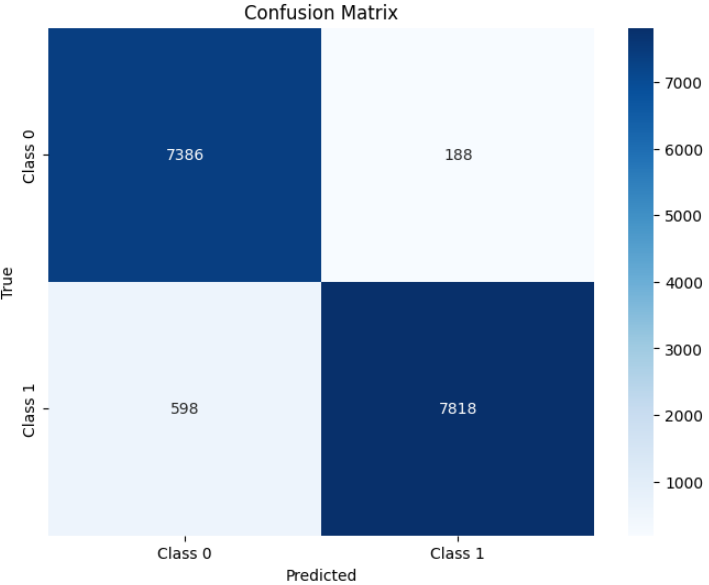
*Hình 4.1: Một mẫu dữ liệu bên trái (0: ảnh thật) và bên phải (1:AI tạo)*

**B - Cấu hình thực nghiệm**

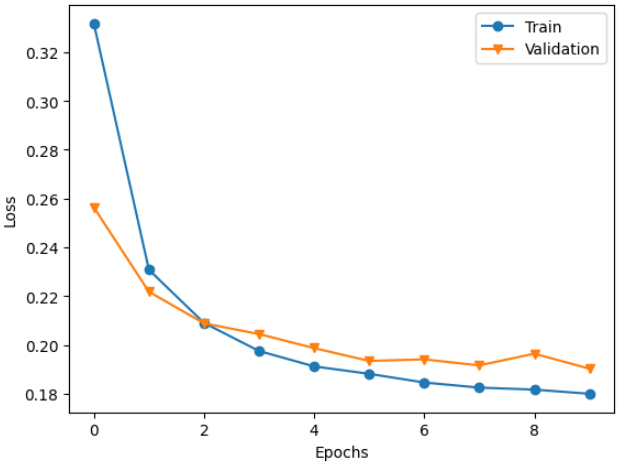
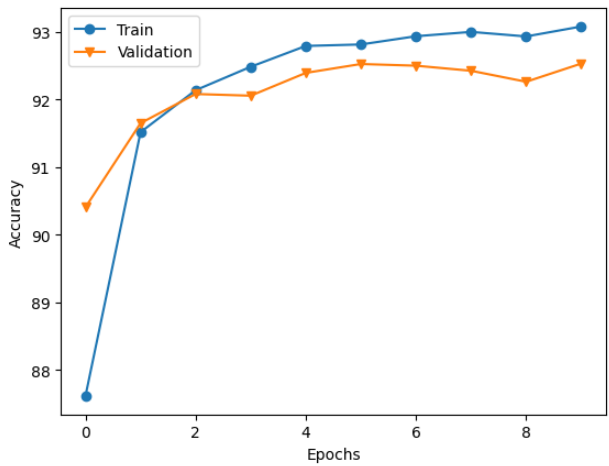
Mô hình được huấn luyện trên GPU P100, Ta thực hiện huấn luyện mô hình bằng bộ tối ưu AdamW với tốc độ học là 1e-4, sử dụng hàm mất mát CrossEntropy với tổng số epoch huấn luyện là 10 epochs. Cả ConvNeXt và ViT đều được fine-tune phần Classifier. Mô hình tốt nhất được lưu dựa trên độ chính xác cao nhất, với các chỉ số đánh giá gồm: Accuracy, Loss và Confusion Matrix.

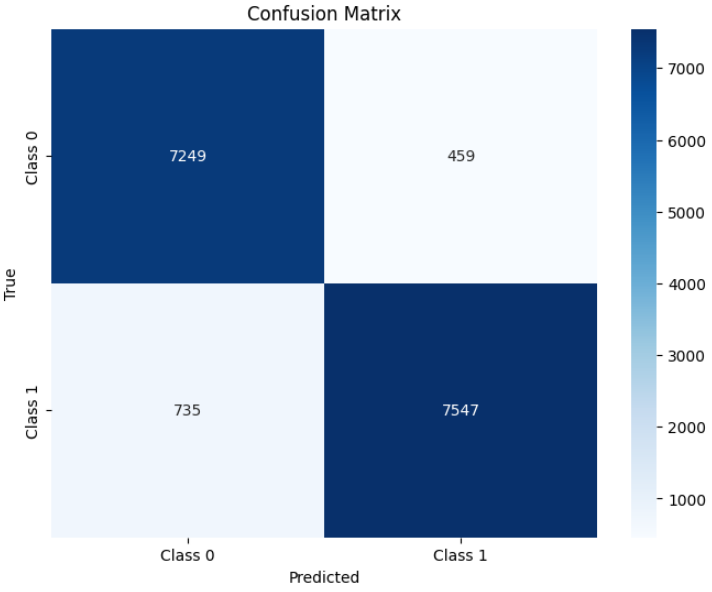
**C - So sánh kết quả**

****

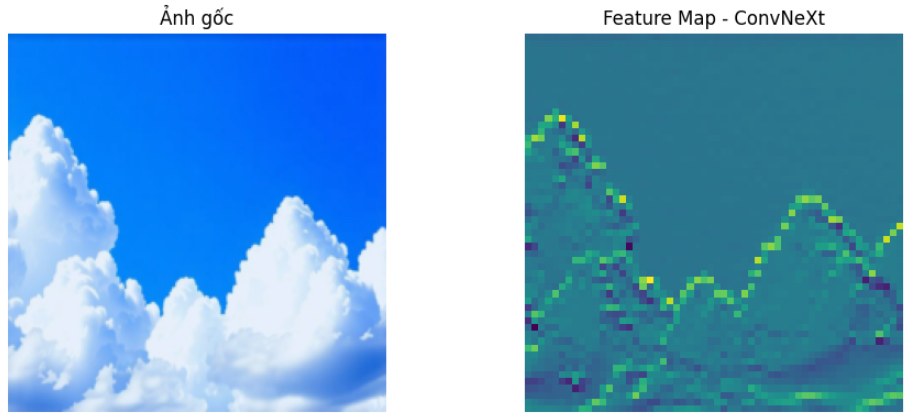
****

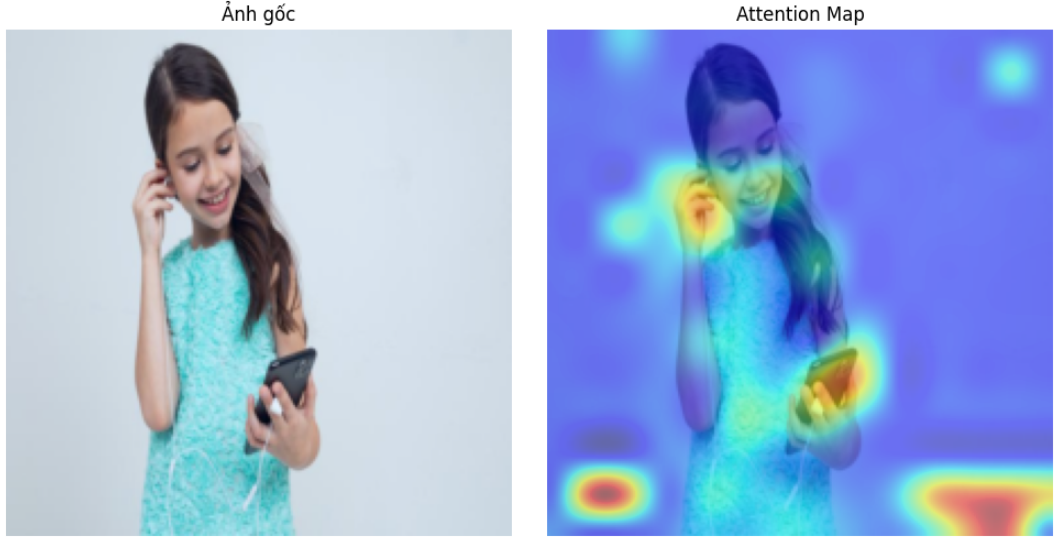
*Hình 4.1: Đánh giá hiệu suất mô hình ConvNeXt thông qua Loss, Accuracy và ma trận nhầm lẫn*

****

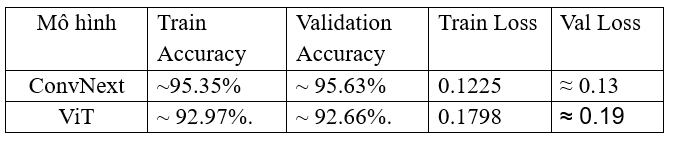
****

*Hình 4.3: Đánh giá hiệu suất mô hình ViT thông qua Loss, Accuracy và ma trận nhầm lẫn*





*Hình 4.2: Đặc trưng của ảnh sau khi đi qua mô hình ConvNeXt và ViT*



*Hình 4.3: Bảng Benchmark của 2 mô hình*

Kết quả thực nghiệm cho thấy **ConvNeXt đạt hiệu suất vượt trội hơn ViT**, thể hiện rõ qua cả độ chính xác và giá trị hàm mất mát. Cụ thể, ConvNeXt đạt độ chính xác kiểm định lên tới 95.63%, cao hơn gần 3% so với ViT, đồng thời có hàm mất mát thấp hơn và ổn định hơn trong suốt quá trình huấn luyện. Điều này chứng tỏ ConvNeXt có khả năng học đặc trưng hiệu quả và khả năng tổng quát hóa tốt hơn.

## **V. KẾT LUẬN**

Trong bài báo này, nhóm tác giả đã trình bày một phương pháp mới để phân loại hình ảnh do AI tạo và hình ảnh do người thật tạo, với hiệu quả được minh chứng qua kết quả thực nghiệm.Cụ thể, hai mô hình chính được thử nghiệm là ConvNext và ViT. Mô hình ConvNext đạt độ chính xác trên tập huấn luyện khoảng 95.35% và trên tập kiểm định khoảng 95.63%. Trong khi đó, mô hình ViT đạt độ chính xác trên tập huấn luyện khoảng 92.97% và trên tập kiểm định khoảng 92.66%.Nhóm tác giả đã tập trung phân tích ba yếu tố quan trọng: dữ liệu, kiến trúc mô hình và chiến lược khởi tạo mô hình, từ đó xây dựng một giải pháp tổng thể có khả năng khái quát tốt trên các bộ dữ liệu mới. Ngoài ra, việc lựa chọn tập dữ liệu huấn luyện một cách hợp lý và sử dụng các mô hình đã được huấn luyện trước ở các tác vụ liên quan, đã giúp cải thiện tính ổn định và hiệu suất tổng thể của hệ thống phân loại.

## **TÀI LIỆU THAM KHẢO**

**[1]** A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with CLIP latents,” *arXiv preprint arXiv:2204.06125*, 2022.

**[2]** R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.

**[3]** Y. Mirsky and T. Mahler, “The creation and detection of deepfakes: A survey,” *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, 2021.

**[4]** S. Wang, X. Yu, and J. Li, “AI-generated image detection based on texture and frequency domain analysis,” *J. Vis. Commun. Image Represent.*, vol. 89, p. 103750, 2023.

**[5]** A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

**[6]** Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “ConvNeXt: Revisiting ConvNets for Image Recognition,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12076–12086.

**[7]** Y. Zhang, P. Sun, H. Qi, and Y. Ma, “Detecting GAN-generated fake images using co-occurrence matrices,” *arXiv preprint arXiv:1903.06836*, 2019.

**[8]** F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Detection of GAN-generated fake images over social networks,” in *Proc. IEEE Conf. Multimedia Information Processing and Retrieval (MIPR)*, 2018.

**[9]** [The architecture of ConvNeXt-Tiny. | Download Scientific Diagram](https://www.researchgate.net/figure/The-architecture-of-ConvNeXt-Tiny_fig3_372021045)

**[10]** [Model Architecture Vision Transformer](https://towardsdatascience.com/is-this-the-end-for-convolutional-neural-networks-6f944dccc2e9/?fbclid=IwY2xjawKg7I5leHRuA2FlbQIxMABicmlkETFxUWYzODlJOUw0VlFPYnhiAR4CLw9FZUfSF9Yzvoj50X4lQXWDQhOOwe_9Jg5YRhht171B4b60ekrQbjoYng_aem_QhggAK8hSZCKkxSkT04EXQ)