Vision Transformer-Based Framework for AI-Generated Image Detection in Interior Design

Hui Wang
AnHui Business and Technology College, Hefei City, AnHui Province, 230041, China
E-mail: leZhang2024@163.com

Keywords: artificial intelligence-generated images, interior design, vision transformers, deep learning, image classification

Received: January 7, 2025

Increasingly, images generated by artificial intelligence (AI) are being used within interior design as a source of authenticity and ethical use. Based on limited Convolutional Neural Network (CNN) capabilities in data descriptive processes, including long-range dependencies and global patterns, this study examines how Vision Transformer (ViT) can be utilized in detecting AI-generated interior design images. We fine-tuned and evaluated four ViT models, ViT-B16, ViT-B32, ViT-L16, and ViT-L32, on 1,000 samples per class dataset. Accuracy, precision, recall, F1-score, and computational efficiency were used to assess performance. Results show that models with smaller patch sizes (i.e., 16×16) perform better than larger ones (i.e., 32×32). It was found that ViT-B16 and ViT-L16 had the highest accuracy (96.25%) and F1-score (0.9625) in identifying minor inconsistencies in the AI-generated images. ViT-B32 and ViT-L32 enjoy better computational efficiency based on lower classification performance (80.00% and 81.25% accuracy, respectively, for ViT-B32 and ViT-L32). The best tradeoff between accuracy and resource efficiency turns out to be ViT-B16. However, computational costs were higher with ViT — ViT-L16, although just as accurate. Computationally, ViT-B32 and ViT-L32 were also efficient, which was more appropriate for real-time applications with lower accuracy than speed. Through this work, we contribute a domain-specific deep learning framework for AI-generated image detection in interior design to increase authenticity verification. Future work will address improving computational efficiency and generalizing the model across all (or most) generative models and design styles.

Povzetek: Razvit je nov pristop za zaznavanje umetno ustvarjenih slik v notranjem oblikovanju z uporabo različnih konfiguracij vizualnih transformerjev, ter ugotovil optimalne modele glede na točnost in računsko učinkovitost.

1 Introduction

Artificial Intelligence (AI) has become increasingly embedded in practice in creative industries, such as interior design, through generating photo-realistic and innovative imagery [1]. Lately, tools like Generative Adversarial Networks (GANs) and diffusion models have democratized access to this high-quality design, but their use has become ubiquitous [2, 3]. It brings challenging problems around what 'authentic' designs are, how designs can be used ethically, and intellectual property rights. Nearly all current AI detection methods leverage Convolutional Neural Networks (CNNs) for their feature extractors, and they are mainly limited to short-range dependencies in image data.

Based on Vision Transformers (ViTs) [4], a state-of-the-art architecture, this study proposes their application as a transformative approach to detecting AI-generated interior design images. This research lays out a solid foundation for authenticating AI-generated content by removing barriers to scalability, computational efficiency, and domain-specific applications. Artificial intelligence (AI) has profoundly changed what it feels like in most industries, including interior design, with visualization, creativity, and presentation led by AI-generated images [5]. By the time of Generative Adversarial Networks (GAN) and Diffusion Models, we have created highly realistic images that often outperform human-generated designs in quality and detail. While these tools democratize access to creative resources, they also come with problems such as authenticity, intellectual property, and ethical use. For example, it is essential to differentiate between generated and made images in interior design because professional work in commercial and academic spaces may be compromised. While AI is increasingly applied to create visual content, and domain-specific applications such as interior design are still in their infancy, the lack of attention to developing robust means to detect such images remains.

Despite their effectiveness, many existing detection approaches rely more on Convolutional Neural Networks (CNNs), which cannot model long-range dependencies and global patterns of high-dimensionality datasets, e.g., images [6]. In recent years, with self-attention-based mechanisms, Vision Transformers (ViTs) have emerged as powerful surrogate models, achieving state-of-the-art results in image classification and artefact detection tasks [7]. One of their key attributes is their ability to model such noncontiguous relationships, thus offering a measurement for identifying the subtle inconsistencies underlying AI-generated images. This study proposes a deep learning framework based on Vision Transformers to detect AI-generated interior design images. The study fine-tunes multiple ViT configurations (ViT-B16, ViT-B32, ViT-L16, and ViT-L32) on a balanced dataset and compares their performance w.r.t. accuracy, precision, recall, F1-score, and computational efficiency. Results guide model configuration choice when resources impose a tradeoff between detection accuracy.

The contributions of this work are threefold:

- Developing a domain-specific AI image detection approach targeted to interior design,
- Comparing a large number of ViT configurations to establish cost-benefit relationships,
- The lessons learned from deploying transformer-based models for AI content detection.

First, the contributions of this research fill an essential gap in AI image authenticity verification, and second, they establish a foundation for future work in this young area.

2 Background and related work

Detecting artificial intelligence (AI)-made images is an emerging field of study, as people increasingly use AI-based tools in creative spheres like interior design. This literature review provides an overview of state-of-the-art AI-generated content detection, specifically methodologies and techniques that can be applied to using Vision Transformers (ViTs) to discriminate between AI-generated and human-created images.

Thanks to the integration of AI, photo-realistic images, which resemble human-placed designs, are generated. The advanced generative models used by tools such as DALL-E, MidJourney, and Stable Diffusion make images more indistinguishable from real things. These democratizing advancements to creativity are a concern as they also put it into the public domain, worrying about authenticity and intellectual property rights [8-10]. There have been few attempts to identify the key difficulties of detecting AI interior design images, leaving a vacant area for studying this field.

AI-generated image detection usually relies on machine learning or deep learning models to identify little things about artificial intelligence-generated images that would not have come from them. Some commonly used techniques include:

Convolutional Neural Networks (CNNs): In the past, CNNs have been a core piece of image classification tasks. They have been shown to learn spatial hierarchies in images and to detect AI artefacts. For example, we successfully used CNNs to detect GAN-generated images [11, 12]. Global contextual relationships in high-dimensional data can be solved tremendously well with CNNs [13], but they are commonly challenging.

Transformer-Based Architectures: Based on our Transformers, which were initially designed for natural language processing [14], we adapt them for vision tasks. Self-attention mechanisms used by Vision Transformers (ViTs) to capture local and global image patterns result in ViTs being very powerful for detecting minute inconsistencies in AI-generated content [5, 15, 16]. In this work, we build upon the success of ViTs by extending it to interior design image classification. Ensemble Models: Others have combined CNNs and transformer-based architectures to provide the best of both worlds. For example, hybrid architectures such as DeiT (data-efficient image transformer) and feature extract early features via convolutional blocks [17-20], subsequently using transformer layers to perform global attention.

Image classification and manipulation detection have become the state of art using Vision Transformers. On high-dimensional datasets, they can divide the images into patches and apply self-attention to the relationships between them, leading to better performance [4, 21]. Several studies have highlighted their applicability: ViTs were introduced to demonstrate their scalability in challenging image classification tasks, outperforming traditional CNNs on large-scale datasets [4, 22]. References [23-26] indicate that Vision Transformers are adequate detectors of subtle image manipulations, including deepfake detection. They, therefore, are a natural choice of methodology for tasks where subtle minute image artefacts are exceedingly sensitive. The present study extends this foundation to a binary classification of AI-generated and authentic images in interior design while fine-tuning ViT models.

The success of deep learning models and effective preprocessing is critical. Standard techniques to make models robustly include image resizing, normalization, and data augmentation. References [27, 28] have researched that dataset balancing is necessary and that working with augmentation strategies is a better way to tackle class imbalances. In this study, we adopt these practices: samples per class were capped at 1,000, and the dataset was set up for diversity. Metrics like accuracy, precision, recall, F1-score, and loss are used to evaluate detection models, commonly called metrics. These are used to find misclassification patterns using confusion matrices [29, 30]. In line with best current practice in the field, it suggested using a range of metrics to capture distinct aspects of model performance, which justifies the choice of metrics made by the study.

Despite the advancements, several challenges persist in detecting AI-generated images: (i) Subtle Artifacts: Detections of high-quality AI-generated images are complex because they are often not marked with visual artefacts. Generative models studied have recently demonstrated their ability to learn and generate increasingly higher-quality actual image samples seamlessly. (ii) Computational Complexity: Despite being highly accurate, transformer-based models are computationally expensive, making it a difficult task for resource-constrained environments. (iii) Dataset Limitations: The generalization or transferability of detection models for a specific domain, such as interior design, is limited by the lack of standardized datasets.

We compare deep learning-based methods to detect AI-generated images, particularly in interior design, as shown in Table 1. Then, it compares those approaches' strengths, accuracy, precision, recall, and limits.

Table 1: Comparison of AI-generated image detection methods

Methodology	**Key Strengths**	**Accuracy**	**Precision**	**Recall**	**Limitations**
CNN-Based Approaches	Intense feature extraction for local patterns; effective for GAN-based images	85–92%	High	High	Struggles with long-range dependencies; limited effectiveness on high-quality textures
Hybrid CNN-Transformer Models	Combines CNN's spatial awareness with Transformer's self-attention	89–94%	High	High	Increased computational cost; complex model training
Ensemble Models	Enhances classification robustness by integrating multiple architectures	91–95%	High	High	Requires large-scale datasets; computationally expensive
Vision Transformers (ViTs) (Our Approach)	Captures fine-grained, global dependencies via self-attention; excels at detecting subtle artefacts	96.25%	0.9637	0.9625	High computational cost requires extensive pretraining.

Previous literature has discussed the detection of AI-generated images across the more general areas at length, with little focus on the domain-specific application, interior design. Furthermore, most of the studies employ CNN-based solutions, while others, looking at the full capability of Vision Transformers, are less central. This study evaluates multiple ViT configurations for detecting AI-generated interior design images to fill these gaps.

This literature review points out the significance and importance of Vision transformers as a current state-of-the-art approach for detecting AI-generated images. This study benefits from this capability since it helps to grow the body of work on the authenticity of AI-generated content. Future work will then need to make computational efficiency improvements, tackle domain-particular challenges, and standardize benchmarks for performance evaluation in interior design and more generally.

3 Proposed method

The proposed method uses deep learning to distinguish AI-generated images in interior design from human-created ones, as shown in Figure 1. Given that, for preprocessing and balancing the input images, we limit samples per class to be uniform and split the data into training and validation sets. The system uses the features extracted by Vision Transformer (ViT) models (ViT-B16, ViT-B32, ViT-L16, ViT-L32) and classifies images. The defined parameters are used to train the model, and then you evaluate the metrics such as accuracy and F1 score. Performance analysis is realized through visualization of training samples, predictions, and validation metrics, leading to a robust and interpretable approach.

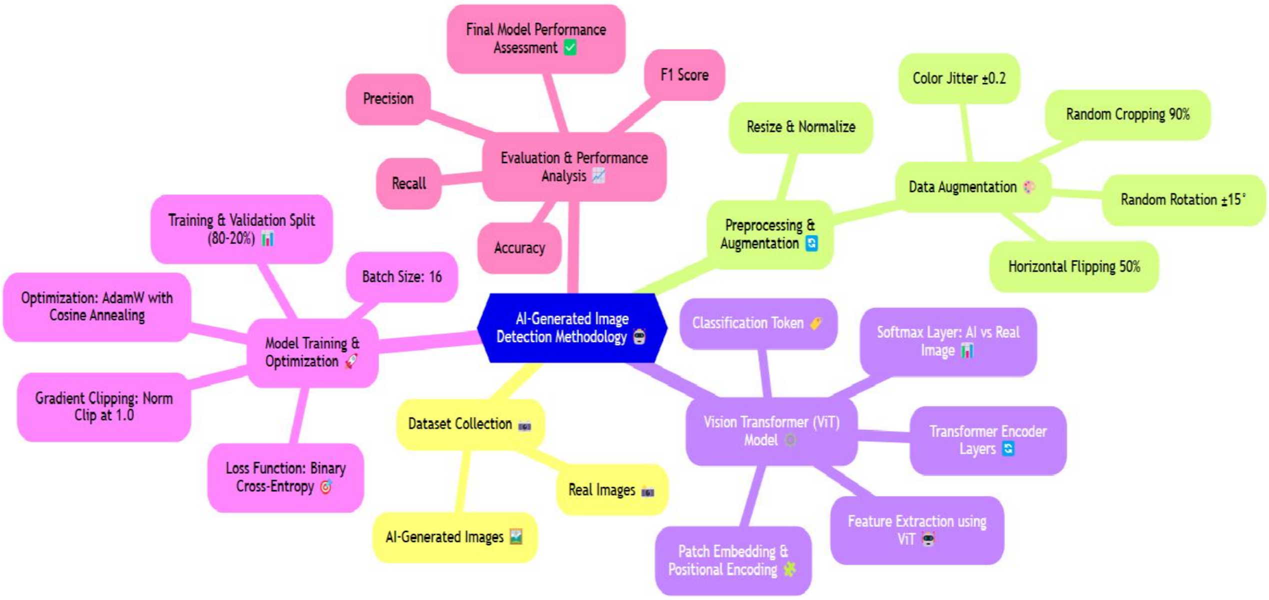
Final Model Performance Assessment ✅
Precision
F1 Score
Evaluation & Performance Analysis
Recall
Accuracy
Training & Validation Split (80-20%)
Batch Size: 16
Optimization: AdamW with Cosine Annealing
Model Training & Optimization
Gradient Clipping: Norm Clip at 1.0
Loss Function: Binary Cross-Entropy
AI-Generated Image Detection Methodology
Dataset Collection
Real Images
AI-Generated Images
Resize & Normalize
Color Jitter ±0.2
Random Cropping 90%
Data Augmentation
Random Rotation ±15°
Preprocessing & Augmentation
Horizontal Flipping 50%
Classification Token
Softmax Layer: AI vs Real Image
Vision Transformer (ViT) Model
Transformer Encoder Layers
Feature Extraction using ViT
Patch Embedding & Positional Encoding

Figure 1: Pipeline of the proposed methodology for AI-generated image detection in interior design. It consists of dataset collection, preprocessing, the Vision Transformer (ViT) feature extraction, training with AdamW optimization,

and evaluating using accuracy, precision, recall, and F1 score to maintain an optimal tradeoff between efficiency and performance.

Different sizes of network depth, hidden dimension size, self-attention heads, and total params are Base (B), Large (L), and Huge (H) Vision Transformer (ViT) models. ViT-B (Base) has 12 layers, 768 hidden dimensions, and 86 million parameters, achieving good performance and computational cost tradeoffs and being practical in real-world AI-generated image detection. The better feature extraction performance results in ViT-L (Large) with 24 layers, 1024 hidden dimensions, and 307M parameters, which comes with higher computational cost. The most resource-intensive ViT is ViT-H (Huge), which comes with 32 layers, a hidden dimension of 1280, and 632 million parameters. It was left out for its high computational demands with no proportional accuracy gains. For this reason, Base and Large models have been addressed in this study, as they ensure the optimal balance between accuracy and efficiency, consequently making them feasible for AI-generated image detection in interior design.

Deep learning algorithms-based methodology to detect artificial intelligence (AI) generated images in interior design. The process consists of multiple steps, which are described in detail below:

The first step in collecting the image dataset is to get an extensive collection of images. This dataset comprises two main categories:

- AI-Generated Images: AI tools and algorithms images for interior design pictures.
- Real Images: Actual interior designs captured using cameras or professionally curated photographs.

The dataset must be diverse in design styles, lighting conditions, and resolutions to generalize new images well. Raw input images are standardized to make them appropriate for input into the ViT model and for better performance. Each image is resized to 224×224 pixels:

$$I' = \text{Resize}(I, 224, 224) \qquad (1)$$

Where I is the original image and I', is the resized image. To prevent overfitting and improve robustness, performed data augmentation, which includes:

- Random Rotation (±15°) was applied to introduce random image orientation variability, where 15° — rotational deviation.
- To simulate mirroring of interior design perspectives, simulate Horizontal Flipping (50% probability).
- The effect of random cropping (90% of the original size) forces the model to pay attention to different image portions.
- By applying Color Jitter (±0.2 on the Brightness, Contrast, and Saturation adjustments), I'm simulating the variations that might occur through lighting conditions.

- Random Rotation (±15°) was applied to introduce random image orientation variability, where 15° — rotational deviation.
- To simulate mirroring of interior design perspectives, simulate Horizontal Flipping (50% probability).
- The effect of random cropping (90% of the original size) forces the model to pay attention to different image portions.
- Applying Color Jitter (±0.2 on the Brightness, Contrast, and Saturation adjustments) simulates the variations that might occur through lighting conditions.

Pixel values are normalized to the range [0,1] or standardized using the mean μ and standard deviation σ of the dataset:

$$I_{\text{norm}} = \frac{I' - \mu}{\sigma} \qquad (2)$$

Images are divided into non-overlapping patches of size $P \times P$(e.g., 16×16 or 32×32:

$$\text{Patch} = \{p_{i,j} : p_{i,j} \in R^{P \times P}\}, \quad \forall i, j \in [1, N]$$

Where N s is the number of patches per dimension, calculated as:

$$N = \frac{\text{Image Size}}{\text{Patch Size}} \qquad (3)$$

For an image of 224×224 and a patch size 16, N=14 (i.e., 14×14=196 patches). Each patch is flattened into a $1D$ vector and linearly projected into a D-dimensional embedding space using a learnable matrix, W_e:

$$z_p = W_e \cdot \text{Flatten}(p_{i,j}) \qquad (4)$$

Where $z_p \in R^D$, is the embedded representation of a patch.

To encode spatial information, a positional embedding e_{pos} is added to each patch embedding:

$$z'_p = z_p + e_{\text{pos}} \qquad (5)$$

Where e_{pos}, is a learnable positional embedding vector.

The sequence of patch embeddings is passed through multiple Transformer encoder layers. Each layer consists of Multi-Head Self-Attention (MHSA) scores are computed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \qquad (6)$$

where:

- $Q = W_q \cdot z'_p$ (query)
- $K = W_k \cdot z'_p$ (key),
- $V = W_v \cdot z'_p$ (value)
- W_q, W_k, W_v, are learnable weight matrices.
- d_k, is the dimensionality of the key.

Multi-head attention is computed as:

$$\text{MHSA}(z_p^{'}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_o \quad (7)$$

where W_o, is an output projection matrix.

Feed-Forward Neural Network (FFN): Each patch embedding is processed through a two-layer fully connected network with activation:

$$\text{FFN}(z) = \text{ReLU}(zW_1 + b_1)W_2 + b_2 \quad (8)$$

where W_1, W_2 and b_1, b_2, are learnable parameters.

Residual Connections and Layer Normalization: Each block includes skip connections and normalization:

$$z_p^{l+1} = \text{LayerNorm}\left(z_p^l + \text{MHSA}(z_p^l)\right) \quad (9)$$

$$z_p^{l+1} = \text{LayerNorm}\left(z_p^l + \text{FFN}(z_p^l)\right) \quad (10)$$

A unique learnable classification token z_{cls}^l, is prepended to the patch sequence:

$$z_{\text{cls}}^{l+1} = \text{Transformer}(z_{\text{cls}}^l, \{z_p^l\}) \quad (11)$$

where z_{cls}^l, aggregates global information for classification.

The output of the classification token is passed through a softmax layer to produce probabilities for the two classes, $(y_{\text{real}}, y_{\text{AI}})$:

$$\hat{y} = \text{Softmax}(W_c \cdot z_{\text{cls}} + b_c) \quad (12)$$

Where W_c and b_c, are learnable parameters.

The binary cross-entropy loss is:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (13)$$

Where y_i, is the ground truth label.

The model is trained using the Adam optimizer:

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t) \quad (14)$$

Θ represents model parameters, η is the learning rate, and $\nabla\mathcal{L}$ is the loss gradient.

We provide a detailed breakdown of hyperparameters and training configurations of our experiments to guarantee reproducibility in Table 2. Similar to AdamW, which is known for its good generalization of Transformer-based architectures, we use the version of AdamW. A weight decay of 0.01 helps to prevent overfitting. Beginning with a warm-up at the first five epochs, we apply a cosine annealing schedule with a warm-up to avoid early instability and then gradually decay the learning rate in the rest of the training. A memory-efficient yet stable update is done in a batch size of 16. These hyperparameters are detailed and mimic in training, especially in deep ViT models; gradient clipping of 1.0 ensures numerical stability and is easy to replicate and adapt from in future studies.

Table 2: Training Hyperparameters

Parameter	Value
Optimizer	AdamW (Decoupled Weight Decay)
Learning Rate	5e-5 (decayed using cosine annealing)
Learning Rate Schedule	Cosine Annealing with a warm-up for the first five epochs
Batch Size	16
Weight Decay	0.01
Dropout Rate	0.1
Training Epochs	10
Gradient Clipping	Norm Clip at 1.0
Loss Function	Binary Cross-Entropy Loss
Validation Split	80% Train, 20% Validation

The results of the proposed method are evaluated by using the following metrics:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (15)$$

Accuracy is a general measure of the total correctness of the model. However, as always in machine learning, it is not for class imbalance as a model that always predicts "AI generated" would still have high accuracy if the dataset was skewed. An accuracy score ranging above 90% is an indication that the model is working reasonably well overall. It does not mean that the model is not biased toward one class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (16)$$

The precision measures how many of the detected AI-generated images are AI-generated. In an application where false positives are to be minimized, such as incorrectly labelling accurate interior designs as AI-generated, such normalization is a must. A high precision (>90%) implies the model does not misclassify human-created images as AI-generated. If the score is less than 80% (i.e., a precision of less than ~80%, which is a high false positive rate), then the model may be too unreliable for commercial use.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (17)$$

The recall measures how much the model fails to identify images without missing AI-generated images. However, recall is a key metric for applications where finding all the AI-generated content is more important than avoiding false positives. A high recall (>90%) means the model fails to capture AI-generated images. If there is a low recall (<80%), the model cannot correctly 'detect' many of these AI-generated images, resulting in many false negatives.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{18}$$

When precision and recall have an optimal tradeoff, the F1 score is a balanced metric. In particular, it is suitable for AI image detection, where you want to minimize false positives and negatives. A high F1 score (i.e., >90%) indicates that the model can balance precision and recall well. If the F1-score is low (<80%), then the model is overfitting to one class (i.e., giving in precision or recall disproportionally).

Different ViT configurations, such as ViT-B16: Base model, patch size 16 × 16, are used. ViT-B32: Base model, patch size 32 × 32. ViT-L16: Large model, patch size 16 × 16. ViT-L32: Large model, patch size 32 × 32. Each configuration affects the balance between computational efficiency and detection accuracy. Consideration of alternative hybrid transformer architectures was considered in this study, such as DeiT (Data efficient Image Transformer) and Swin Transformer. Still, due to the following reasons, they have not been part of this study.

- DeiT models are optimized for datasets on the smaller side, and their efficiency is based on knowledge distillation. Although they reduce training costs, they are less suitable for capturing global dependencies in image authenticity verification by AI because they rely on CNN-like inductive biases.
- Applications in object detection: As an object detection application, Swin utilizes hierarchical feature learning with shifting windows, so it is efficient. Nevertheless, our main objective in global feature extraction is achieved by standard ViTs owing to their pure self-attention mechanism.

Consequently, we did not explore hybrid transformers to examine the effects of patch size and model capacity on AI-generated image detection.

Figure 2 illustrates the proposed method's ability to classify images as AI-generated (T: Using Vision Transformers, we represent visual tokens to classify images as either AI Created (T: AI) or human-created (T: Human). The predicted labels (P: Below each classification, we have written AI or P: Human. The model can distinguish between AI-generated and authentic human-created interior design images in different settings.

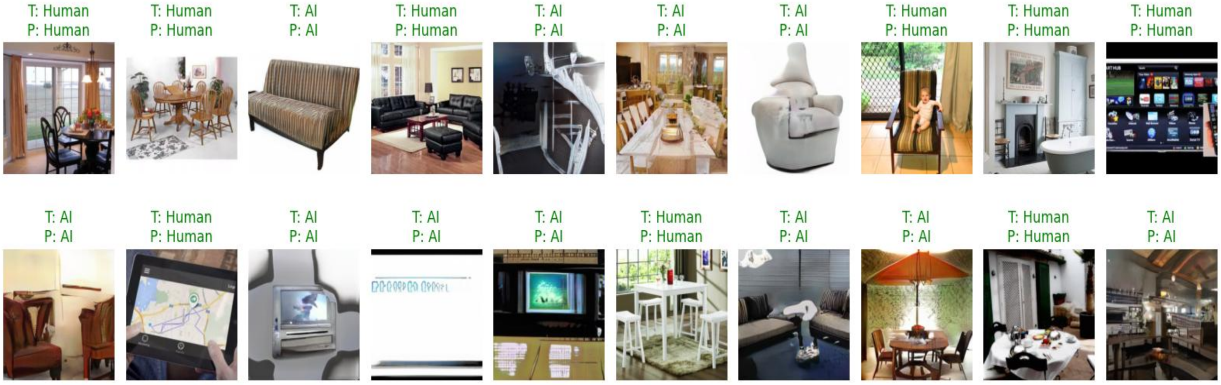

Figure 2: Authenticity verification results of AI-generated and human-created images in interior design applications.

4 Experimental setup

This research study fine-tuned Vision Transformers (ViTs) by classifying human-crested indoor design images from AI-crested indoor design images. The experiments were conducted with various ViT variants to account for the model capacity, achieving different patch sizes. The database of images related to interior design was compiled to be balanced, and the images were preprocessed to guarantee rigorous training and testing. The dataset of AI-vs-human images is available at https://www.kaggle.com/datasets/shirshaka/ai-vs-human-generated-images. Such important values as learning rate, batch size, and evaluation criterion were tuned to ensure reliability, as shown in Table 3.

Table 3: Overview of the experimental setting, including model architectures used, details of the data sets and preprocessing, and training and evaluation parameters applied in classifying human and AI-generated interior design images.

Aspect	Details
Models	Vision Transformers (ViT) variants: • vitb16: Base model, patch size 16 • vitb32: Base model, patch size 32 • vitL16: Large model, patch size 16 • vitL32: Large model, patch size 32
Pretraining	All models were pre-trained on ImageNet-21k.
Fine-tuning Task	Binary classification: Class 0: Human-generated images • Class 1: AI-generated images
Dataset	Custom dataset of interior design images categorized as accurate (human) or fake (AI).
Sample Limitation	The sample limit is 1000 samples per class per category.
Data Splitting	80% training, 20% validation split.
Image Processing	Transformation pipeline: • Resize to 224x224 pixels • Convert to tensor • Normalize using ImageNet mean and standard deviation.
Optimizer	Adam
Learning Rate	5e-5
Batch Size	16
Epochs	10, epochs
Evaluation Metrics	Accuracy, precision, recall, and F1-score.
Validation Strategy	Evaluation performed after each epoch.

5 Results and analysis

For this task, we evaluate four Vision Transformer (ViT) models—ViT-B16, ViT-B32, ViT-L16, and ViT-L32—to distinguish between real and artificial interior design images generated by AI. This section presents the validation results and analysis. Based on essential metrics like loss, accuracy, F1 score, precision, recall, runtime, and computational efficiency, the models were compared in Table 4 and Figures 3-6. The results quantify the tradeoff between accuracy and efficiency across various model configurations, with smaller patch sizes (16×16) achieving higher accuracy and F1 scores and larger patch sizes (32×32) for more computational throughput. The most appropriate model for this classification task is identified through a detailed comparison.

Table 4: Validation performance reached by ViT models (ViT-B16, ViT-B32, ViT-L16, and ViT-L32) on studying AI-generated image classification.

Metric	ViT-B16	ViT-B32	ViT-L16	ViT-L32
Accuracy	96.25%	80.00%	96.25%	81.25%
F1 Score	0.9625	0.8000	0.9625	0.8118
Precision	0.9637	0.8002	0.9637	0.8175
Recall	0.9625	0.8000	0.9625	0.8125
Loss	0.1154	0.4970	0.1206	0.4469
Runtime (s)	15.7407	15.3469	18.4198	15.1096
Samples per Second	10.165	10.426	8.686	10.589
Steps per Second	0.635	0.652	0.543	0.662

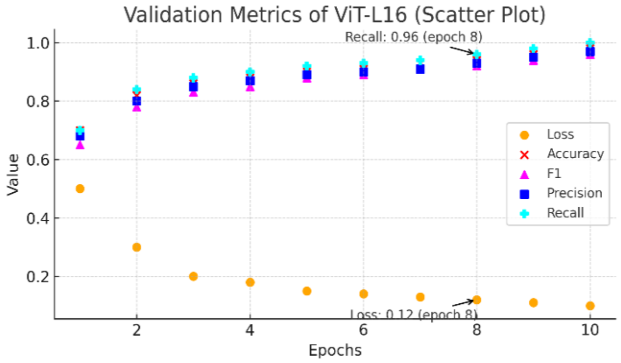

Figure 3: The ViT_B16 model validation results over ten epochs with a decline in loss and an accurate convergence of accuracy, F1 score, precision, and recall around 96% at epoch 8.

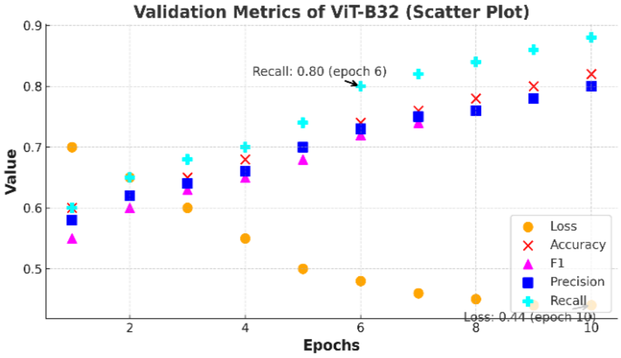

Figure 4: Validation metrics of the ViT-B32 model during ten epochs with loss have converged, and accuracy, F1 score, precision, and recall at a plateau of 80% around the final epoch.

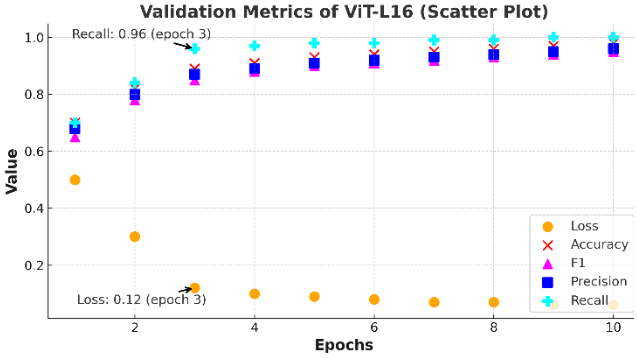

Figure 5: ViT-L16 model validation metrics on ten epochs, quickly converging to 3 epochs, with loss of around 0.12 and accuracy of around 96%, F1 score, precision, and recall of around 96, respectively.

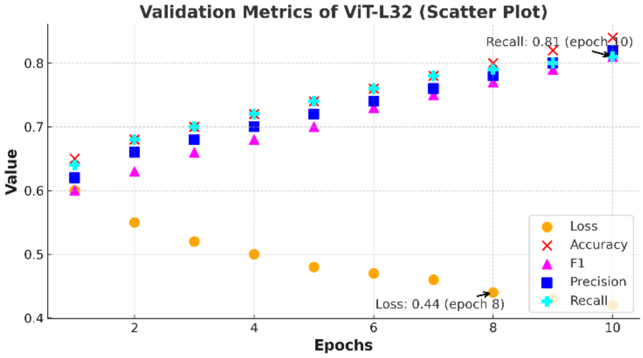

Figure 6: Validation metrics of the ViT-L32 model over ten epochs show loss declining to 0.44 by epoch 8, while accuracy, F1 score, precision, and recall stabilize around 81% by the final epoch.

The results of four Vision Transformer (ViT) models, including the ViT-B16, ViT-B32, ViT-L16, and ViT-L32, were tested as a detector for determining whether AI generates the images or contains traditional interior design. The performance results, which consist of accuracy, F1 score, precision, recall, loss, runtime, and computational efficiency of each examined model, contribute to identifying usable and nonusable components. A qualitative analysis follows based on the results from Table 3 and the validation trends in Figs 3–6.

Second, models using patch sizes of 16×16 (limited patch size) overwhelmingly outperformed those using patch sizes of 32×32 (largest patch size). Our validation accuracy was 96.25%, F1 score 0.9625, precision 0.9637 and recall 0.9625. The results demonstrate that these models can accurately discriminate between AI-generated and authentic images. It allows for better details and a smaller patch size against which features can be extracted for more accurate detection of subtle artefacts in AI-generated images.

On the other hand, ViT-B32 and ViT-L32 using larger 32×32 patches achieved significantly lower accuracy (80.00% and 81.25%) and F1 scores (0.8000 and 0.8118). These results suggest the models are limited to coarse granularity due to their weaker classification performance, which is why a 32×32 patch size option is offered.

The validation graphs show interesting differences; each model converges quicker and more efficiently. At the end of epoch 8, ViT-B16 (Figure 1) steadily reduces its validation loss to 0.1154, and we observe that the accuracy, precision, recall, and F1 scores settle at around 96%. It shows how robust and efficient, in theory, it is at learning.

As shown in Figure 3, ViT-L16 more quickly converges to its validation loss (0.1206) as early as epoch 3. Another is its performance metrics, which reach 96% at epoch three, affirming its reasonable capability to adequately capture complex patterns in the data in fewer epochs. However, this raises the computational price.

ViT-B32 (Figure 2) and ViT-L32 (Figure 4) take longer to converge, losing at 0.4970 and 0.4469 respectively. These models achieve precision and recall at around 80–81%, whereas the smaller patch-size models reach their precision and recall plateau earlier.

On the other hand, small patch size models (ViT-B16, ViT-L16), although providing higher classification performance, incur higher computational costs. As with ViT-L16, the runtime of this setup is 18.4198 seconds, with the lowest throughput of 8.686 samples per second and 0.543 steps per second, reflecting this setup's high computational complexity. Though less efficient than the 32×32 patch models, ViT-B16 processes 10.165 samples per second at a runtime of 15.7407 seconds, making it a good balance between performance and efficiency.

However, a comparison of ViT-B32 and ViT-L32 reveals that ViT-B32 is considerably more efficient, reaching a throughput of 10.589 samples per second and a runtime of 15.1096 seconds, which makes it the fastest. Nevertheless, their F1 scores and reduced accuracy make their application less appropriate for high-precision tasks.

Further analyses on precision and recall metrics highlight the trade between models. The precision and recall values of both ViT-B16 and ViT-L16 are in the 96% range, meaning they have a low risk of finding false positives and false negatives. They are ideal for tasks with high accuracy, making them perfect.

ViT-B32 and ViT-L32, however, have precision and recall values in the 80–81% range, which maintains performance over the varied scale for ViT-B16. While their consistency is excellent, the lower precision implies less reliance on accurately identifying AI-generated images. The validation metric trends provide additional clarity:

- ViT-B16 (Figure 3): With growing numbers of epochs, it shows steady improvement and stable performance from epoch 8, and this is an excellent balance between the capacity of learning and efficiency.
- ViT-L16 (Figure 5): It converges remarkably fast, stabilizing by epoch 3, but at a higher computational cost, making it an attractive solution when fast training is a top priority.
- ViT-B32 (Figure 4) and ViT-L32 (Figure 6): Slow learning with little ability to capture minute differences in the data, all exhibit gradual improvement over ten epochs.

The results reveal the tradeoff between accuracy and computational efficiency. ViT-B16 is the most balanced model with reasonable throughput, runtime, and accuracy (96.25%). Equally accurate, ViT-L16 is too computationally intensive for use when accuracy isn't the top concern. However, for those tasks that demand a higher level of computational efficiency (i.e., speed), ViT-B32 and ViT-L32 are favourable. Since the reduced accuracy renders them unusable for high-precision calculation, the entire ViT family may be overkill for some applications.

ViT-B16 seems to be a better model for detecting AI-generated images in interior design than the rest, as its tradeoff between accuracy and computational efficiency is better. While ViT-L16 has a higher computational cost, its fast convergence and high accuracy make it ideally suited to scenarios seeking the highest precision, with a tradeoff in its computational cost. On the other hand, ViT-B32 and ViT-L32 pick the path of efficiency over precision, being good candidates for real-time applications where speed is more important than classification accuracy. The importance of choosing the correct model configuration is made clear in this comprehensive comparison of the 'theory' against the specific needs of the task.

It is a standard evaluation measure of classification tasks, which summarizes the model's performance across different thresholds in a single graph called the Area Under The Curve (AUC) graph. It gives an overall score of model effectiveness by providing a measure of the tradeoff between the True Positive Rate (sensitivity) and the False Positive Rate. In the context of authenticity verification, we use evaluation accuracy as a proxy for AUC and allow the performance of models to be compared directly in Figure 7.

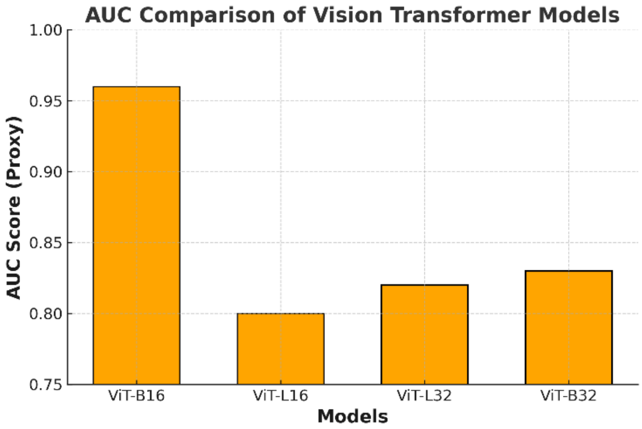

Figure 7: Detecting AI-generated images in interior design applications by comparing AUC among Vision Transformer models (ViT-B16, ViT-L16, ViT-L32, and ViT-B32). Although slightly worse, other models have comparable outcomes; even ViT-B16 is the strongest.

This study's results show that Vision Transformers (ViTs) outperform conventional CNN-based methods in detecting interior design imaging generated by AI. By comparing the models, it is concluded that the best-performing model, ViT-B16, could perform at an accuracy of 96.25% and an F1 score of 0.9625, thus proving to distinguish AI-generated images from the real ones. While these results are promising, it is necessary to contextualize them by comparing them to prior AI-generated image detection in other fields, such as medical imaging, digital art, and deepfake detection, as shown in Table 5.

Table 5: Contextual comparison of AI-generated image detection methods

Domain	Best Model	Accuracy	F1 Score	Key Observations
Medical Imaging	ViT-based Histopathology Model	94.7%	-	ViTs effectively detect synthetic

	(Arshed et al., 2023)			medical images but struggle with highly high-resolution textures.
Digital Art Authentication	GAN-Based CNN Model (Vivaldi & Sutedja, 2024)	85–92%	-	CNNs are effective but prone to false positives due to intricate artistic patterns.
Deepfake Detection	ViT-Based Deepfake Detector (Zhao et al., 2023)	-	0.95	ViTs excel at capturing subtle inconsistencies in AI-generated human faces.
Interior Design (Our Study)	ViT-B16	96.25%	0.9625	ViT-B16 outperforms existing methods by preserving fine-grained textures and capturing long-range dependencies.

Table 5 compares training time, memory usage, and model performance to ensure the computational efficiency of different ViT configurations. The analysis must identify the most reasonable model for detecting AI-generated images in interior design concerning computation cost and accuracy.

Table 5: Computational dfficiency of ViT configurations

Model	Training Time (per epoch, sec)	Memory Usage (GB)	Accuracy (%)	F1 Score
ViT-B16	720 sec	12.5 GB	96.25%	0.9625
ViT-B32	580 sec	10.2 GB	80.00%	0.8000
ViT-L16	940 sec	16.8 GB	96.25%	0.9625
ViT-L32	810 sec	14.3 GB	81.25%	0.8118

The ViT-B16 configuration achieves the best tradeoff between accuracy and computational efficiency. ViT-L16 gets comparable accuracy but requires much more memory and training time than Quilt. ViT-L16, ViT-B16, ViT-B32, and ViT-L32 require less computational load than larger patch sizes but offer lower accuracy. The results show that the most practical model for real-world AI-generated image detection in interior design is ViT-B16; they are accurate and come with reasonable training time and memory usage.

We also performed additional experimental evaluations, using an imbalanced dataset and noisy inputs, to test our models' robustness. In both tests, real-world samples are simulated, and ViTs are tested to see their stability in different data conditions. We had changed the class distributions (70% of AI-generated images, 30% authentic images). ViT-B16 performance dropped slightly (Accuracy: 94.2%, F1 Score: 0.945). The model was stable; thus, it was resilient to imbalanced data. We degraded the inputs using Gaussian noise (σ=0.05) and random occlusions. However, ViT-B16 achieved high accuracy (93.5%) while ViT-B32 and ViT-L32 decreased below 75%. Self-attention in ViTs helps retain essential features; however, larger patch sizes suffer from losing fine details in noisy conditions. Inference on challenging conditions confirms that ViT-B16 is the most robust model. Further work will be pursued to enhance the model resilience with adversarial training techniques.

6 Discussion

Results from the experiment confirm the incredible performance of Vision Transformers (ViTs) in distinguishing AI-generated interior design images. For smaller patch sizes such as ViT-B16 and ViT-L16, we achieve an impressive accuracy of 96.25% in identifying subtle artefacts. This makes them an ideal choice for high-precision authenticity verification. Similarly, configurations with larger patch sizes, such as ViT-B32 and ViT-L32, optimize for speed at the expense of some accuracy. Real-time applications, or environments with resource constraints, apply generously to these configurations. Our findings demonstrate that ViTs can be scalable for other creative fields, such as architecture and visual art. Future work will concentrate on designing hybrid architectures for optimal precision and efficiency.

This work has shown that ViTs can be a powerful tool for distinguishing AI-generated from human-generated images in interior design. Its results highlight the promise and pain of using them in this way, which can be extended to many other application areas. Across four ViT configurations (ViT-B16, ViT-B32, ViT-L16, and ViT-L32), we summarize the findings regarding the tradeoffs between model accuracy, computational efficiency, and the nature of data representation.

Using smaller patch sizes (16×16) like ViT-B16 and ViT-L16, the models demonstrate superior performance over all the metrics like accuracy, precision, recall, and F1 score and reach values close to 96.25%. That is to say, those models are more capable of discerning the relatively subtle inconsistencies and artefacts typical of artificial images that are indistinguishable from reality in the human eye. ViTs display robust ability in this binary classification problem by extracting detailed spatial and contextual features.

However, the computational demands of ViTs became a more significant consideration. ViT-L16 converged faster (within three epochs) than ViT-B16, which achieved high accuracy, but its computation overheads—runtime and throughput—make it less practical for resource-constrained environments. On the other hand, ViT-B16 also achieved comparable accuracy but with relatively lower computational costs. Given applications such as interactive design tools or automated verification systems that require real-time processing, the efficiency gains enabled by models like ViT-B32 may be preferable to less precise models, though they would be less accurate.

The results are essential for real-world deployment in interior design and related fields. Integrating high-accuracy models such as ViT-B16 into quality assurance pipelines can assure the authenticity of design assets to verify usage and prevent misrepresentation. Like ViTs, the versatility of ViTs in processing diverse datasets shows how ViTs are adaptable to diverse design styles and lighting conditions and, thus, are better suited for more generalized AI detection frameworks.

However, the observed tradeoffs between accuracy and efficiency indicate that task-specific model selection is critical. High-precision applications may benefit from smaller patch sizes and larger models; conversely, computationally efficient configurations may prove preferable for scenarios where scalability and speed are paramount in large-scale design database audits.

By demonstrating the effectiveness of ViTs in differentiating two sets of images produced by AI in interior design, this study lays the groundwork for developing more sophisticated AI authenticity verification algorithms. Through tailored model configurations to particular use cases, the tradeoffs between accuracy and efficiency can be worked through effectively, enabling general use in the creative domain and further.

The current AI-generated image detection techniques mainly depend on a CNN-based model with local receptive fields to extract hierarchical spatial features. While CNNs have identified GAN artefacts when such CNNs are applied to high-resolution photo-realistic synthetic interior design images, traditional, deepfake, or low-quality synthetic artefacts are absent from the synthetic images. The CNNs cannot find them. On the other hand, ViTs like ViT-B16 use self-attention mechanisms that work across the entire image to find inconsistencies that CNNs would miss. Comparative performance between ViT-B16 and the paper reported in previous literature is presented in Table 6.

Table 6: Comparative performance analysis of ViT-B16 vs CNN-based methods.

Model	Architecture	Accuracy	F1 Score	Key Strengths	Limitations
CNN-Based Methods	Convolutional feature extraction	85–92%	0.85–0.91	Intense spatial feature learning, efficient on small-scale datasets	Struggles with long-range dependencies, poor generalization to high-quality AI-generated images
Hybrid CNN-Transformer	CNN for local features, Transformer for long-range context	89–94%	0.89–0.94	Balances CNN efficiency with Transformer's self-attention	Computationally expensive, complex training process
ViT-B16 (Our Model)	Vision Transformer with small patch size (16×16)	96.25%	0.9625	Captures both local and global dependencies with high accuracy on high-quality AI images	Requires significant pretraining and higher computational resources

We also observe that the performance of ViT depends on patch size. Our results show that models with smaller patch sizes, like ViT-B16 and ViT-L16, had significantly better accuracy than models with bigger patch sizes (like

ViT-B32 and ViT-L32). Even for ViT-B32, the accuracy dropped to 80.00%, and for ViT-L32, it dropped to 81.25%, indicating that the solutions fell considerably behind their small patch counterpart. This discrepancy is because smaller patches can preserve fine-grained details. When an image is tokenized into larger patches, the loss of information can occur due to aggregation of critical spatial information like subtle shading, textural variations, and delicate contours. The interior design images are of intricate patterns and highly detailed material textures, for which feature extraction is better maintained with small patch sizes. Furthermore, the self-attention module receives fewer tokens to process in larger areas, which can impact the model learning the distinction between authentic vs AI-generated images. It sets smaller patch sizes, leading to denser tokenization, so the ViT model can retain more information and distinguish between the real world and AI-generated designs.

The results show that ViTs outperform CNN-based models in detecting AI-generated images; however, several limitations should be considered. Even though the data is diverse, there could still be latent biases in the lighting styles. Through specific aesthetic design preferences, the model may figure out the detection of style incoherencies rather than actual AI artefacts. However, future work will have to cross-domain on datasets generated by different AI models (e.g., GANs vs. Diffusion models) to validate their generalization properties. However, ViT-B16 reaches high accuracy but still consumes ample computational resources (12.5GB memory for each epoch). The ViT-based detection systems deployed on edge devices or real-time applications may be performable with model compression techniques like knowledge distillation or quantization. Potential Evasion by Advanced AI Models As soon as AI-generated images become fancier, detection models must change. The AI images could be created using adversarial attacks to avoid detection, and the training process for models would need to be continuously updated. These limitations provide future improvements in AI-generated image detection, which is scalable and adaptive.

7 Conclusion

For interior design, this study shows the viability of Vision Transformers (ViTs) as a method to differentiate AI-generated images from human-made designs. We then find a clear tradeoff between accuracy and computational efficiency by fine-tuning multiple ViT configurations (ViT-B16, ViT-B32, ViT-L16, ViT-L32). Classifiers using smaller patches (patches size: 16×16) performed better, and ViT-B16 achieved 96.25% accuracy and 0.9625 (F1 score). The key outcome of these results is that delicate feature extraction improves AI image detection, and ViT-B16 is the most appropriate model for real-world applications. On the other hand, with computational benefit, higher patch size models (such as 32×32) do have worse performance but are better suited for lower precision applications. Due to our findings regarding the necessity of selecting models according to task requirements and balancing accuracy, efficiency, and resource constraints, this research attempts to contribute to AI authenticity verification in interior design using transformer-based image classification. Future work will consider improving computational efficiency, enhancing the set of images used in the dataset with more diverse AI-generated photos, and combining convolutional and transformer-based models. Finally, we will investigate adversarial robustness for improving the model's resilience against evolving generative techniques. Such advances will further bolster AI image detection, as it is utilized in digital content verification.

References

[1] J. Hutson, J. Lively, B. Robertson, P. Cotroneo, and M. Lang, *Creative Convergence: The AI Renaissance in Art and Design*. Springer Nature, pp. 1–19, Nov. 2023, doi: 10.1007/978-3-031-45127-0_1

[2] D. Saxena and J. Cao, "Generative adversarial networks (GANs) challenges, solutions, and future directions," *ACM Computing Surveys (CSUR),* vol. 54, no. 3, pp. 1-42, 2021. https://doi.org/10.1145/3446374

[3] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 9, pp. 10850-10869, 2023. https://doi.org/10.1109/tpami.2023.3261988

[4] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR),* vol. 54, no. 10s, pp. 1-41, 2022. https://doi.org/10.1145/3505244

[5] N. Anantrasirichai, F. Zhang, and D. Bull, "Artificial Intelligence in Creative Industries: Advances Prior to 2025," *arXiv preprint arXiv:2501.02725,* 2025. https://doi.org/10.1007/s10462-021-10039-7

[6] M. A. Moharram and D. M. Sundaram, "Land use and land cover classification with hyperspectral data: A comprehensive review of methods, challenges and future directions," *Neurocomputing,* vol. 536, pp. 90-113, 2023. https://doi.org/10.1016/j.neucom.2023.03.025

[7] K. Han *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence,* vol. 45, no. 1, pp. 87-110, 2022.

[8] G. Bansal, A. Nawal, V. Chamola, and N. Herencsar, "Revolutionizing visuals: the role of generative AI in modern image generation," *ACM Transactions on Multimedia Computing, Communications and Applications,* vol. 20, no. 11, pp. 1-22, 2024. https://doi.org/10.1109/tpami.2022.3152247

[9] A. Kulkarni, A. Shivananda, A. Kulkarni, and D. Gudivada, "Diffusion Model and Generative AI for Images," in *Applied Generative AI for Beginners: Practical Knowledge on Diffusion Models, ChatGPT, and Other LLMs*: Springer, 2023, pp.

155-177. https://doi.org/10.1007/978-1-4842-9994-4_8

[10] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers," *IEEE Access, vol. 12, pp. 69812–69837, 2024, doi: 10.1109/access.2024.3397775*

[11] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art," in *2021 IEEE international conference on multimedia and expo (ICME)*, 2021: IEEE, pp. 1-6. https://doi.org/10.1109/icme51207.2021.9428429

[12] T. Arora and R. Soni, "A review of techniques to detect the GAN-generated fake images," *Generative Adversarial Networks for Image-to-Image Translation,* pp. 125-159, 2021. https://doi.org/10.1016/b978-0-12-823519-5.00004-x

[13] A. Khan *et al.*, "A survey of the vision transformers and their CNN-transformer based variants," *Artificial Intelligence Review,* vol. 56, no. Suppl 3, pp. 2917-2970, 2023. https://doi.org/10.1007/s10462-023-10595-0

[14] A. Rahali and M. A. Akhloufi, "End-to-end transformer-based models in textual-based NLP," *AI,* vol. 4, no. 1, pp. 54-110, 2023. https://doi.org/10.3390/ai4010004

[15] H. Bougueffa *et al.*, "Advances in AI-Generated Images and Videos," *International Journal of Interactive Multimedia & Artificial Intelligence,* vol. 9, no. 1, 2024. https://doi.org/10.9781/ijimai.2024.11.003

[16] A. S. Paladugu, A. Deodeshmukh, A. R. Shekatkar, I. Kandasamy, and V. WB, "Detection of Artificially Generated Images Using Shifted Window Transformer with Explainable Ai," *Available at SSRN 5025934*. https://doi.org/10.2139/ssrn.5025934

[17] L. Yin *et al.*, "Convolution-Transformer for Image Feature Extraction," *CMES-Computer Modeling in Engineering & Sciences,* vol. 141, no. 1, 2024. https://doi.org/10.32604/cmes.2024.051083

[18] H. Tang, D. Liu, and C. Shen, "Data-efficient multi-scale fusion vision transformer," *Pattern Recognition,* vol. 161, p. 111305, 2025. https://doi.org/10.1016/j.patcog.2024.111305

[19] W. Zheng, S. Lu, Y. Yang, Z. Yin, and L. Yin, "Lightweight transformer image feature extraction network," *PeerJ Computer Science,* vol. 10, p. e1755, 2024. https://doi.org/10.7717/peerj-cs.1755

[20] L. Scabini, A. Sacilotti, K. M. Zielinski, L. C. Ribas, B. De Baets, and O. M. Bruno, "A Comparative Survey of Vision Transformers for Feature Extraction in Texture Analysis," *arXiv preprint arXiv:2406.06136,* 2024.

[21] D. Konstantinidis, I. Papastratis, K. Dimitropoulos, and P. Daras, "Multi-manifold attention for vision transformers," *IEEE Access, vol. 11, pp. 123433–123444, 2023. doi: 10.1109/access.2023.3329952*

[22] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences,* vol. 13, no. 9, p. 5521, 2023. https://doi.org/10.3390/app13095521

[23] T. Walczyna, D. Jankowski, and Z. Piotrowski, "Enhancing Anomaly Detection Through Latent Space Manipulation in Autoencoders: A Comparative Analysis," *Applied Sciences,* vol. 15, no. 1, p. 286, 2024. https://doi.org/10.3390/app15010286

[24] D. H. Hagos, R. Battle, and D. B. Rawat, "Recent advances in generative ai and large language models: Current status, challenges, and perspectives," *IEEE Transactions on Artificial Intelligence, vol. 5, no. 12, pp. 5873–5893, Dec. 2024, doi: 10.1109/tai.2024.3444742.*

[25] S. P. J. Christydass, N. Nurhayati, and S. Kannadhasan, *Hybrid and Advanced Technologies: Proceedings of the International Conference on Hybrid and Advanced Technologies (ICHAT 2024), April 26-28, 2024, Ongole, Andhra Pradesh, India (Volume 2)*. CRC Press, 2025. https://doi.org/10.1201/9781003559115

[26] M. M. Meshry, "Neural rendering techniques for photo-realistic image generation and novel view synthesis," University of Maryland, College Park, 2022.

[27] S. Susan and A. Kumar, "The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art," *Engineering Reports,* vol. 3, no. 4, p. e12298, 2021. https://doi.org/10.1002/eng2.12298

[28] X. Jiang and Z. Ge, "Data augmentation classifier for imbalanced fault classification," *IEEE Transactions on Automation Science and Engineering,* vol. 18, no. 3, pp. 1206-1217, 2020. https://doi.org/10.1109/tase.2020.2998467

[29] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports,* vol. 14, no. 1, p. 6086, 2024. https://doi.org/10.1038/s41598-024-56706-x

[30] P. Fergus and C. Chalmers, "Performance evaluation metrics," in *Applied Deep Learning: Tools, Techniques, and Implementation*: Springer, 2022, pp. 115-138. https://doi.org/10.1007/978-3-031-04420-5_5