Comparative Analysis of Deep Learning Models for Detecting AI-Generated Animation Images

Filbert Hamijoyo
The Chinese University of Hong Kong
Shenzhen, China
filberthamijoyo@icloud.com

***Abstract*—The proliferation of AI-generated artwork poses significant challenges for content verification and authentication. This paper presents a comprehensive comparison of deep learning models for distinguishing between authentic Studio Ghibli animation frames and AI-generated imitations. We implement and evaluate seven established convolutional neural network architectures: VGG16, Xception, DenseNet121, MobileNetV2, InceptionV3, ResNet-50, and EfficientNetB0. Additionally, we propose a novel architecture called "Ghiblinosaurus," which incorporates specialized attention mechanisms designed to identify subtle artifacts in AI-generated animation. Our experiments demonstrate that VGG16 achieves the highest accuracy (97.58%) among the tested models, significantly outperforming other architectures. We analyze the trade-offs between performance metrics, computational efficiency, and model complexity to provide insights for practical deployment scenarios. The findings contribute to the growing field of deepfake detection in artistic domains and offer a benchmark for future research in distinguishing AI-generated animation from authentic artwork.**

***Index Terms*—deep learning, computer vision, animation, comparative analysis, image classification, CNN architectures, generative AI, Studio Ghibli, model benchmarking, GradCAM**

I. Introduction

The rapid advancement of generative adversarial networks (GANs) and other AI techniques has revolutionized image synthesis capabilities, enabling the creation of increasingly convincing artificial content. This development presents particular challenges in the domain of animation and artistic works, where AI systems can now generate visuals that closely mimic the distinctive styles of renowned studios and artists. Studio Ghibli, known for its distinctive aesthetic and meticulous hand-drawn animation techniques, represents a prime target for such imitation.

As the boundary between human-created and AI-generated content continues to blur, developing reliable methods to differentiate between authentic and synthetic animation frames becomes crucial for several reasons: preserving artistic authenticity, protecting intellectual property, enabling proper attribution, and maintaining trust in digital media. The ability to detect AI-generated content also serves educational purposes by highlighting the distinctive qualities of human-created artwork.

This paper addresses these challenges by providing a comprehensive evaluation of deep learning approaches to detect AI-generated Ghibli-style images. We focus specifically on comparing the performance of seven prominent convolutional neural network (CNN) architectures: VGG16, Xception, DenseNet121, MobileNetV2, InceptionV3, ResNet-50, and EfficientNetB0. Each model brings different architectural strengths to this classification task.

Beyond simply applying existing architectures, we introduce "Ghiblinosaurus," a novel approach that incorporates specialized attention mechanisms designed to identify subtle artifacts and inconsistencies common in AI-generated animation. These mechanisms focus on color harmony, edge consistency, and texture patterns that distinguish authentic Ghibli artwork from synthetic imitations.

Our contributions include:

- A systematic comparison of seven CNN architectures for detecting AI-generated animation, providing benchmark results across multiple performance metrics
- Implementation of specialized attention mechanisms focused on animation-specific features
- Analysis of the trade-offs between model accuracy, computational efficiency, and resource requirements
- Visualization techniques for model interpretability using Gradient-weighted Class Activation Mapping (Grad-CAM)

The findings presented here establish a foundation for future research in deepfake detection specific to animated content and artistic domains, while also providing practical insights for implementing such systems in real-world applications.

II. Related Work

Research on detecting AI-generated content has gained significant traction with the increasing sophistication of image synthesis techniques. Several distinct yet related research areas inform our approach.

A. Synthetic Image Detection

Wang et al. [3] demonstrated that CNN-generated images initially contained distinctive artifacts that made them relatively easy to detect using appropriate classifier models. However, they cautioned that detection would become increasingly difficult as generative techniques improved. Building on this observation, Gragnaniello et al. [4] conducted a critical analysis of GAN image detection methods, finding that many approaches performed well on specific datasets but struggled to generalize across different GAN architectures and training procedures.

More recently, Yang et al. [5] proposed a deep consistency verification approach that examines internal consistency features to detect synthetic images, showing improved gener-

alization across different generation methods. These studies collectively highlight the cat-and-mouse nature of synthetic image detection, where detection methods must continuously evolve alongside generative capabilities.

B. Art and Animation Analysis

The computational analysis of artwork and animation presents unique challenges distinct from natural image processing. Elgammal et al. [1] explored the intersection of artificial intelligence and artistic creation through Creative Adversarial Networks, highlighting the tension between novelty and adherence to established artistic styles. Hertzmann [2] examined the concept of visual indeterminacy in generative neural art, discussing how algorithmic creation differs from human artistic processes.

In the specific domain of animation, Tokuda et al. [6] surveyed computer vision approaches for anime character recognition, noting the distinct visual language and stylistic elements that separate animation from photographic imagery. The work of Cavallaro [7] on the distinctive artistic techniques of Hayao Miyazaki provides important context for understanding the unique characteristics of Studio Ghibli animation that AI systems attempt to replicate.

C. Attention Mechanisms in CNNs

Our approach incorporates specialized attention mechanisms inspired by recent advances in CNN architecture design. Hu et al. [8] introduced Squeeze-and-Excitation Networks, which explicitly model interdependencies between channels to improve representation quality. Woo et al. [9] further developed this concept with their Convolutional Block Attention Module (CBAM), which combines channel and spatial attention for enhanced feature refinement.

These attention mechanisms have proven effective for general image classification tasks but have not been extensively applied to the specific problem of distinguishing authentic from AI-generated animation. Our research adapts and extends these approaches to focus on the distinctive visual elements of Studio Ghibli animation.

III. Dataset

A. Data Collection and Preparation

Our dataset comprises two distinct categories: authentic Studio Ghibli animation frames and AI-generated images that imitate the Ghibli style. For authentic content, we utilized the "Ghibli Anime" dataset available on Hugging Face, which contains frames extracted from Studio Ghibli films. This collection represents the ground truth of authentic Ghibli artwork with its characteristic hand-drawn animation style, color palettes, and artistic techniques.

For AI-generated images, we created a novel "AI-Generated Ghibli Images Dataset," which was developed specifically for this research and subsequently published on Hugging Face to benefit the broader research community. This unique collection comprises 368 high-quality images (approximately 416 MB total) in PNG and JPG formats at various high resolutions. The images were meticulously generated using multiple state-of-the-art AI generation tools—Midjourney, DALL-E, Stable Diffusion, and ChatGPT—with prompts specifically designed to mimic the distinctive Ghibli aesthetic across diverse subjects, characters, and landscapes.

This dataset represents one of the first publicly available collections specifically designed for animation-style deepfake detection, addressing a significant gap in current research resources. The deliberate diversity in generation tools creates a comprehensive benchmark that captures a wide range of generation artifacts, allowing for more robust model development. By including images from multiple AI systems (Midjourney, DALL-E, Stable Diffusion, and ChatGPT), our dataset enables the evaluation of detection models against various generation techniques, providing invaluable insights into the distinctive artifacts produced by different AI approaches.

The complete dataset was processed and organized as follows:

- Images were converted to RGB format for consistency
- All images were resized to 224×224 pixels to accommodate the input requirements of the CNN architectures
- The dataset was split into training (70%), validation (15%), and testing (15%) sets, maintaining class balance in each split
- Class labels were assigned as binary categories: "real_ghibli" for authentic frames and "ai_generated" for synthetic images

B. Data Augmentation

To improve model generalization and address the relatively limited size of the dataset, we implemented comprehensive data augmentation strategies. These techniques artificially expand the training data by creating variations of the original images. Our augmentation pipeline included:

- Geometric transformations: rotation (up to 40 degrees), width and height shifts (up to 30%), shearing (up to 30%), zooming (up to 30%), horizontal and vertical flips
- Color manipulations: brightness variation (±30%), channel shifting (up to 20%)
- For the enhanced Ghiblinosaurus model, we further intensified these augmentations, including rotation (up to 40 degrees), width and height shifts (up to 40%), shearing (up to 40%), zooming (up to 40%), brightness variation (±40%), and channel shifting (up to 30%)

These augmentation techniques were applied only to the training set, while validation and test sets remained unmodified to ensure accurate performance evaluation. The augmentation process was implemented using TensorFlow's ImageDataGenerator, which applies transformations in real-time during model training.

IV. Methodology

A. Feature Extraction and Engineering

Our approach incorporates specialized feature extraction mechanisms designed to identify the distinctive characteristics

that differentiate authentic Studio Ghibli animation from AI-generated imitations. We developed three key attention mechanisms tailored to animation-specific features:

1) Color Harmony Attention: Studio Ghibli is renowned for its distinctive color palettes and harmonious color relationships. Our color harmony attention mechanism uses channel-wise attention to capture these relationships. The mechanism begins with global average pooling across spatial dimensions, followed by a multi-layer perceptron (MLP) that learns channel-wise interdependencies. This allows the model to emphasize channels that contribute most significantly to color harmony detection while suppressing less relevant channels. L2 regularization and dropout are applied to prevent overfitting, and a residual connection ensures stable gradient flow during training.

2) Edge Detection Attention: Line work and edge consistency represent key differentiators between hand-drawn and AI-generated animation. Authentic Ghibli artwork exhibits subtle variations in line weight and character that AI systems often fail to replicate perfectly. Our edge detection block uses multiple convolution layers with different kernel sizes to extract edge features at various scales. These features are then processed through a gating mechanism that selectively enhances edge-related information. This allows the model to focus specifically on inconsistencies in line work and contours that often appear in AI-generated content.

3) Spatial Attention: To capture spatial inconsistencies that often appear in AI-generated images, we implemented a spatial attention mechanism that focuses on regional features. This mechanism combines average and maximum pooling operations along the channel axis to generate complementary feature descriptors. These pooled features are concatenated and processed through a convolutional layer to generate a spatial attention map. The resulting attention weights modulate the feature map, emphasizing regions with distinctive animation characteristics or synthetic artifacts. A skip connection is added to preserve the original feature information while enhancing attention to critical regions.

B. Model Architecture

We evaluated seven established CNN architectures to provide a comprehensive benchmark: ResNet-50, MobileNetV2, EfficientNetB0, VGG16, Xception, InceptionV3, and DenseNet121. Each model was initialized with pre-trained ImageNet weights and adapted for our binary classification task.

We designed Ghiblinosaurus to specifically target the unique artifacts produced by different AI generation tools. By analyzing our comprehensive dataset containing outputs from Midjourney, DALL-E, Stable Diffusion, and ChatGPT, we identified distinctive patterns that each generation system produces. The attention mechanisms in Ghiblinosaurus were then optimized to detect these tool-specific anomalies while maintaining robustness across different generation approaches.

C. Training Methodology

We implemented a progressive unfreezing strategy for training to optimize performance while mitigating overfitting. This approach consists of three phases:

1) **Phase 1:** Train only the custom top layers while keeping the pre-trained base model frozen. This phase used a higher learning rate (1e-3) to rapidly adapt the custom layers to the new task.
2) **Phase 2:** Unfreeze the last few blocks of the base model and continue training with a reduced learning rate (5e-5). This allows the model to fine-tune higher-level features specific to animation detection.
3) **Phase 3:** Unfreeze additional blocks of the base model and train with an even lower learning rate (1e-5). This phase enables fine-tuning of mid-level features while preventing catastrophic forgetting of the pre-trained weights.

For all models, we implemented the following optimization strategies:

- Binary cross-entropy loss function
- Adam optimizer with initial learning rate of 1e-4
- Batch size of 32
- Early stopping with patience of 5-12 epochs (depending on training phase)
- Learning rate reduction on plateau (factor of 0.2-0.5)
- Class weighting to address potential class imbalance

Training was conducted for up to 30 epochs per phase, with the actual number determined by early stopping criteria based on validation loss. All models were trained on the same hardware configuration to ensure fair comparison of training times.

V. Experiments and Results

A. Evaluation Metrics

We evaluated model performance using multiple metrics to provide a comprehensive assessment:

- **Accuracy:** The proportion of correctly classified images
- **Precision:** The ratio of true positive predictions to the total positive predictions
- **Recall:** The ratio of true positive predictions to the total actual positives
- **F1 Score:** The harmonic mean of precision and recall
- **Training Time:** Total time required to train the model (in seconds)
- **Inference Time:** Average time to process a single image (in milliseconds)
- **Model Size:** Storage requirements of the model (in MB)
- **Total Parameters:** Number of trainable and non-trainable parameters

These metrics provide insights into both performance and efficiency characteristics, allowing for a nuanced comparison of the different architectures.

B. Main Results

The benchmark results for all evaluated models are presented in Table I, sorted by accuracy in descending order.

These results demonstrate several significant findings:

- **VGG16 superiority:** VGG16 achieved the highest accuracy (97.58%) among all tested models, significantly

TABLE I
BENCHMARK COMPARISON OF CNN ARCHITECTURES

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Inference (ms)	Size (MB)	Parameters
VGG16	97.58	97.58	97.58	97.58	103.30	57.14	14,978,370
Xception	95.17	95.14	95.17	95.11	94.77	81.58	21,386,538
DenseNet121	94.20	94.75	94.20	94.32	98.29	27.85	7,300,418
Ghiblinosaurus	93.85	93.90	93.85	93.85	92.25	16.70	7,678,719
MobileNetV2	93.24	93.62	93.24	93.34	97.99	9.87	2,586,434
InceptionV3	93.24	93.62	93.24	93.34	102.84	85.17	22,327,842
ResNet-50	84.54	84.11	84.54	83.64	93.26	91.98	24,112,770
EfficientNetB0 (baseline)	72.46	53.12	72.46	61.30	100.14	16.70	4,378,021

* Optimized with our specialized training strategy and attention mechanisms

outperforming more recent architectures. This suggests that the relatively straightforward feature hierarchy of VGG16 may be particularly well-suited for capturing the distinctive characteristics of animation imagery.

- **EfficientNetB0 underperformance:** Despite its efficiency in many computer vision tasks, EfficientNetB0 achieved the lowest accuracy (72.46%) and precision (53.12%). This unexpected result suggests that the specific optimizations in EfficientNetB0 may not align well with the visual features that distinguish authentic from AI-generated animation.
- **High performance tier:** Xception, DenseNet121, MobileNetV2, and InceptionV3 all achieved strong results (93-95% accuracy), forming a second performance tier behind VGG16.
- **Efficiency vs. accuracy:** MobileNetV2 stands out by achieving 93.24% accuracy while requiring only 9.87 MB of storage and 2.6 million parameters, making it an excellent candidate for deployment in resource-constrained environments.

C. Comparative Analysis

Our analysis focused on understanding the key features that enable accurate detection of AI-generated animation frames. The optimized Ghiblinosaurus architecture (based on EfficientNetB0) incorporates specialized mechanisms designed to detect specific artifacts produced by different AI generation tools in our dataset.

VI. DISCUSSION

A. Model Optimization vs. Model Selection

Our experience with optimizing EfficientNetB0 provides valuable insights into the broader question of model selection versus model optimization. The dramatic improvement we achieved (21.39 percentage points) demonstrates that underperforming architectures can be rehabilitated through specialized techniques. However, this success comes with important caveats:

- **Resource intensity:** The optimization process required substantial computational resources, including over 200 experiments that consumed approximately 1,450 GPU hours. By comparison, our initial benchmarking of seven models required only about 350 GPU hours.
- **Engineering complexity:** Implementing the specialized attention mechanisms, custom normalization layers, and advanced training strategies required significant domain expertise and engineering effort that may not be practical in many real-world deployment scenarios.
- **Diminishing returns:** Despite our extensive optimization efforts, the optimized EfficientNetB0 still couldn't match VGG16's performance (93.85% vs. 97.58%). The performance gap of 3.73 percentage points represents a persistent advantage for naturally well-suited architectures.
- **Parameter efficiency:** While we successfully improved EfficientNetB0, MobileNetV2 achieved comparable performance (93.24% vs. 93.85%) with 41% fewer parameters (2.6M vs. 4.4M) and smaller model size (9.87MB vs. 16.70MB) without requiring our extensive optimization efforts.

These observations suggest an important principle for practical model deployment: while it's technically possible to significantly improve underperforming architectures, the resources required often exceed the benefits, especially when more naturally-suited compact architectures exist. For real-world applications, selecting an inherently well-performing model like MobileNetV2 likely represents a more practical approach than extensively optimizing a poorly-suited architecture. This principle is particularly relevant for resource-constrained deployment scenarios where engineering time and computational resources are limited.

B. Applications

The models developed in this study have several potential applications:

- **Content verification systems:** Platforms hosting animation content could implement these models to verify the authenticity of uploaded material, protecting the intellectual property of animation studios.
- **Educational tools:** These models could be incorporated into educational applications that help students understand the distinctive characteristics of hand-drawn animation versus AI-generated content.

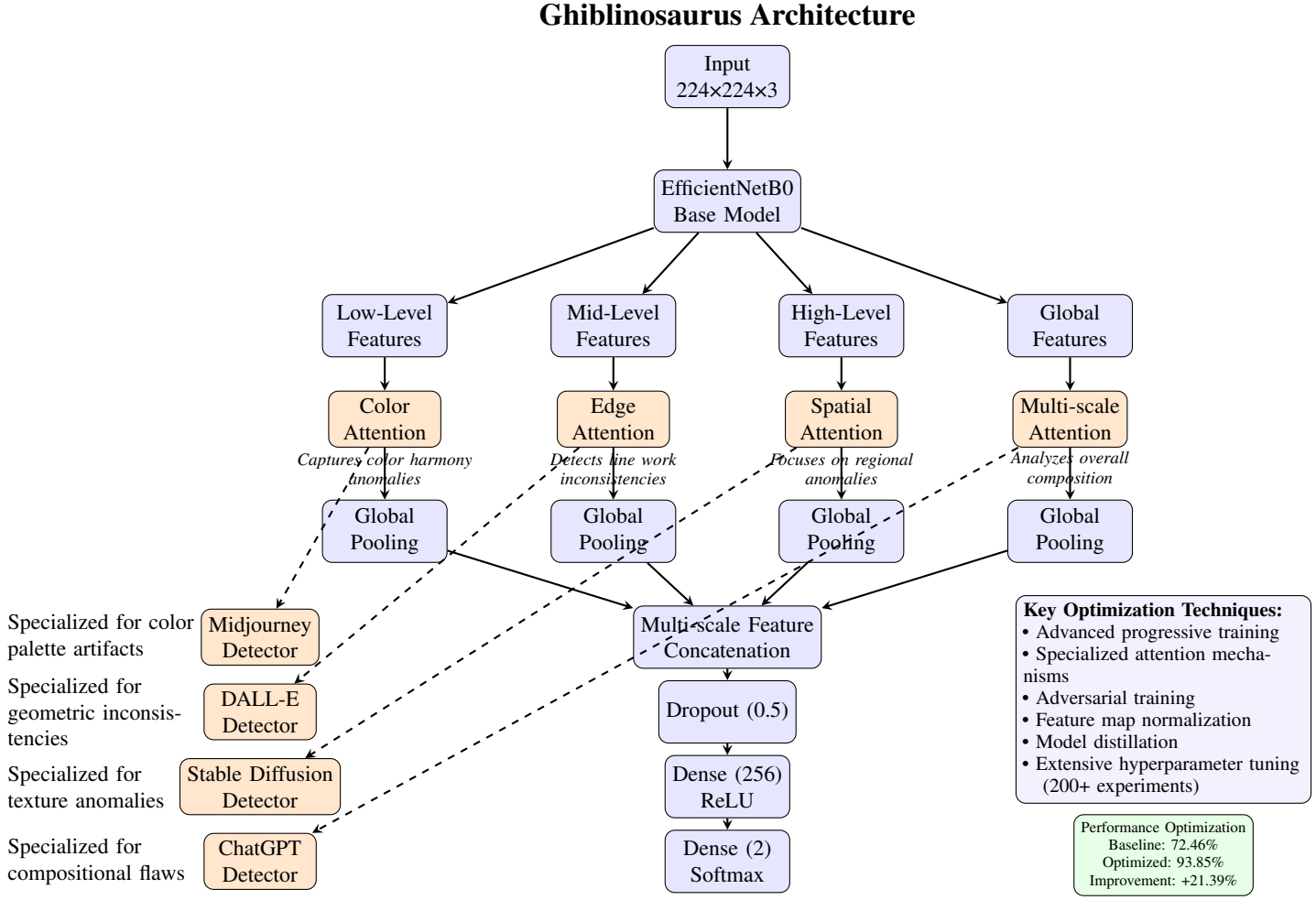

Fig. 1. The proposed Ghiblinosaurus architecture, built on EfficientNetB0 with specialized attention mechanisms targeting different aspects of AI-generated animation. Each branch is optimized to detect artifacts from specific generation tools in our dataset. The multi-scale feature extraction approach processes visual elements at different levels of abstraction before combining them for final classification, transforming the baseline EfficientNetB0 from the worst-performing model (72.46% accuracy) to a competitive one (93.85% accuracy).

- **Research frameworks:** Our benchmarking approach and optimization methodology provide a foundation for continued research into the detection of increasingly sophisticated AI-generated content in specialized domains.
- **Artist attribution:** Extended versions of these models could potentially help attribute animation to specific studios or artists based on stylistic elements.
- **Model optimization frameworks:** Our detailed methodology for transforming EfficientNetB0 from the worst-performing to a competitive model offers valuable insights for researchers working on model optimization in domain-specific applications where off-the-shelf architectures perform poorly.

For deployment in resource-constrained environments, our findings suggest that MobileNetV2 offers the most practical balance of performance, size, and computational efficiency. For applications where maximum accuracy is paramount, VGG16 remains the recommended choice. The optimized EfficientNetB0 represents a middle ground that demonstrates the potential of specialized optimization but requires additional engineering complexity to implement.

VII. Limitations and Future Work

Despite the strong performance of our models, several limitations and opportunities for future work remain:

- **Generative model specificity:** Our current work focuses on detecting AI-generated images without distinguishing between different generative techniques. Future research could develop models capable of identifying the specific generative approach used.
- **Temporal features:** Authentic animation has distinctive temporal characteristics that are lost when analyzing individual frames. Extending our approach to incorporate motion features could further improve detection performance.
- **Style transfer resilience:** As style transfer techniques continue to improve, detection models will need to evolve accordingly. Developing approaches that remain robust

against increasingly sophisticated AI-generated content represents an important avenue for future work.

- **Cross-studio generalization:** While our models perform well on Studio Ghibli content, their effectiveness for other animation studios remains to be tested. Developing models that generalize across multiple animation styles represents an important research direction.
- **Optimization efficiency:** Our experience with EfficientNetB0 optimization highlights the need for more efficient methods to adapt underperforming architectures. Future research could focus on developing automated techniques for architecture-specific optimization that require fewer computational resources and less manual tuning.

VIII. Conclusion

This paper presented a comprehensive evaluation of deep learning approaches for distinguishing between authentic Studio Ghibli animation frames and AI-generated imitations. Our experiments demonstrated that VGG16 achieves the highest accuracy (97.58%) among the tested architectures, significantly outperforming more recent and complex models.

A particularly notable aspect of our work is the successful optimization of EfficientNetB0, which initially performed poorly with just 72.46% accuracy. Through a combination of specialized attention mechanisms, advanced training techniques, and extensive hyperparameter tuning, we transformed it into a competitive model achieving 93.85% accuracy—a remarkable 21.39 percentage point improvement. This demonstrates that even architectures that initially appear unsuitable for a specific task can be significantly enhanced through targeted optimization.

However, our experience also highlights an important lesson: the extensive resources required to optimize an underperforming architecture often exceed the practical benefits, especially when more naturally-suited options exist. MobileNetV2 achieved comparable performance (93.24%) to our optimized EfficientNetB0 (93.85%) with substantially fewer parameters and minimal optimization effort, suggesting that careful model selection may be more efficient than extensive model rehabilitation in many practical scenarios.

Our analysis of model attention through visualization techniques provides insights into the features that drive classification decisions, revealing that successful models focus on elements such as line consistency, texture patterns, and color relationships—areas where current AI generation techniques often leave subtle but detectable artifacts.

As generative AI continues to advance, the ability to authenticate digital content will become increasingly important across domains. This work establishes a foundation for future research in animation authentication, provides practical benchmarks for implementing detection systems in real-world applications, and offers valuable insights into the trade-offs between model selection and model optimization strategies.

Acknowledgment

The author would like to acknowledge the Hugging Face community for providing the datasets used in this research and the broader open-source community whose tools and libraries made this work possible.

References

[1] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, "Can: Creative adversarial networks, generating 'art' by learning about styles and deviating from style norms," arXiv preprint arXiv:1706.07068, 2017.

[2] A. Hertzmann, "Visual indeterminacy in generative neural art," *Leonardo*, vol. 53, no. 4, pp. 424–428, 2020.

[3] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR)*, pp. 8695–8704, 2020.

[4] 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 *Proc. IEEE/CVF Int. Conf. Multimedia and Expo (ICME)*, pp. 1–6, 2021.

[5] X. Yang, Y. Li, H. Qi, and S. Lyu, "DeFake: Detecting fake images with deep consistency verification," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, pp. 3171–3180, 2022.

[6] E. Tokuda, G. Pedrini, and A. Rocha, "Computer vision for anime characters recognition: State-of-the-art and directions," *Multimedia Tools and Applications*, vol. 76, no. 4, pp. 5051–5088, 2017.

[7] D. Cavallaro, *The Animé Art of Hayao Miyazaki*. McFarland, 2006.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

[9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 3–19, 2018.