# Enhanced Attention-Based Classification of Human-Created vs AI-Generated Anime Images Using MobileNetV2

**Dan Nguyen Vu**
University of Engineering and Technology,
Vietnam National University - Hanoi, Vietnam
23020351@vnu.edu.vn

## Abstract

The proliferation of AI-generated anime images presents significant challenges for online platforms, artists, and consumers. In this paper, we address the problem of distinguishing between human-created and AI-generated anime images using deep learning techniques. We present a novel approach that enhances the MobileNetV2 architecture with channel attention mechanisms and optimized test-time augmentation. Our method achieves 97.28% accuracy and 97.29% F1-score on a balanced dataset of 5,700 images (2,850 human-created, 2,850 AI-generated), surpassing previous state-of-the-art results. The proposed model effectively captures subtle differences between authentic and synthetic anime artwork while maintaining a lightweight architecture suitable for deployment in resource-constrained environments. Our approach demonstrates that strategic architectural enhancements and inference-time techniques can significantly improve performance without increasing model complexity. This work contributes to the growing field of AI-generated content detection and provides valuable tools for protecting the intellectual property of human artists.

## 1 Introduction

> *"If a machine can paint a thousand masterpieces in an hour, yet has never felt the anguish of creation or the joy of inspiration, can its output truly be called art? Does the soul manifest not in the final work, but in the struggle to create it?"*

The rapid advancement of generative artificial intelligence has revolutionized digital content creation, particularly in the domain of artistic imagery. AI-generated anime art, facilitated by models such as Stable Diffusion, DALL-E, and specialized systems like NovelAI, has become increasingly sophisticated and difficult to distinguish from human-created artwork. While this technology offers exciting creative possibilities, it also raises significant concerns for the anime art community.

Human anime artists invest considerable time, effort, and creativity into developing their skills and producing original artwork. The emergence of AI systems capable of generating high-quality anime images in seconds threatens these artists' livelihoods and creative recognition. Furthermore, many AI art generators are trained on datasets containing copyrighted artwork without proper attribution or consent, raising ethical and legal questions about the generation and distribution of such content.

Online art platforms and marketplaces face the challenge of distinguishing between human-created and AI-generated artwork to implement fair policies. Currently, many platforms rely on manual verification, which is time-consuming, subjective, and increasingly difficult as AI-generated images become more convincing. There is an urgent need for automated systems that can accurately identify AI-generated anime images to:

- Protect the intellectual property and economic interests of human artists
- Enable platforms to enforce transparent content policies
- Provide consumers with clarity about the origin of the artwork they view and purchase
- Support attribution and proper crediting of creative work

Recent research by Kusuma *et al.* [9] demonstrated the feasibility of detecting AI-generated anime images using transfer learning with MobileNet architectures, achieving up to 97.2% accuracy. However, their approach used a relatively small dataset (1,000 images) and did not explore attention mechanisms or advanced ensemble techniques that could further improve performance.

In this paper, we present a novel approach to detecting AI-generated anime images that builds upon previous work while introducing several key innovations:

1. A larger, more diverse dataset comprising 5,700 images (2,850 human-created, 2,850 AI-generated) with varying characteristics
2. An enhanced MobileNetV2 architecture incorporating channel attention mechanisms to focus on discriminative features
3. An optimized test-time augmentation strategy that significantly improves classification robustness
4. An ensemble prediction system that combines multiple model variants for better performance

Our approach achieves 97.28% accuracy, surpassing previous state-of-the-art results while maintaining a lightweight architecture suitable for deployment in resource-constrained environments. The proposed model effectively captures subtle differences between authentic and synthetic anime artwork without requiring extensive computational resources.

The remainder of this paper is organized as follows: Section 2 reviews related work in AI-generated image detection and attention mechanisms in computer vision. Section 3 details our methodology, including the dataset, model architecture, and training approach. Section 4 presents experimental results and comparisons with state-of-the-art methods. Finally, Section 5 concludes the paper and discusses future research directions.

## 2 Related Work

### 2.1 AI-Generated Image Detection

The detection of AI-generated or manipulated images has become an active research area in response to rapid advancements in generative models. Early approaches focused on detecting GAN-generated images by analyzing artifacts and inconsistencies that were common in first-generation models [12]. Wang *et al.* [21] demonstrated that CNN-based classifiers could effectively distinguish between real and GAN-generated images by learning subtle patterns characteristic of synthesis processes.

As generative models improved, researchers shifted toward more sophisticated detection methods. Liu *et al.* [11] proposed a global texture enhancement approach for detecting fake faces, leveraging the observation that synthetic images often exhibit distinctive texture patterns different from natural images. Their Gram-Net architecture achieved significant improvements over traditional CNN models by explicitly modeling global texture statistics.

In the anime domain, Kusuma *et al.* [9] specifically addressed AI-generated anime detection using transfer learning with MobileNetV2 and MobileNetV3 architectures. Their study used 1,000 anime images (half AI-generated using NovelAI, half from Danbooru2021) and achieved accuracies between 96.8% and 97.2%. Their work established an important baseline for our research but was limited by dataset size and did not explore attention mechanisms or ensemble techniques.

### 2.2 Transfer Learning for Image Classification

Transfer learning has emerged as a powerful paradigm for leveraging knowledge gained from large-scale datasets to improve performance on specialized tasks with limited data [14]. In computer vision,

pre-trained models such as VGG, ResNet, and MobileNet have been widely adapted for various classification tasks through fine-tuning [23].

MobileNetV2 [18], designed for mobile and edge devices, offers an excellent balance between computational efficiency and accuracy. It employs an inverted residual structure with linear bottlenecks, significantly reducing parameters while maintaining performance. This architecture has been successfully applied to numerous classification tasks, including medical image analysis [5], object detection [3], and facial recognition [2].

Several studies have explored optimizing transfer learning for binary classification tasks. Tajbakhsh *et al.* [19] demonstrated that fine-tuning outperforms training from scratch in medical image analysis when data is limited. Kornblith *et al.* [7] showed that models performing better on ImageNet generally transfer better to other tasks, though this correlation is not perfect and can be improved through architectural modifications.

### 2.3 Attention Mechanisms in Computer Vision

Attention mechanisms have significantly enhanced CNN performance by enabling models to focus on the most relevant features for a given task. The pioneering work of Hu *et al.* [4] introduced Squeeze-and-Excitation Networks (SENet), which explicitly model channel interdependencies through a channel attention mechanism. This approach recalibrates channel-wise feature responses adaptively, improving representation power with minimal computational overhead.

Building upon channel attention, Woo *et al.* [22] proposed the Convolutional Block Attention Module (CBAM), which combines both channel and spatial attention. The channel attention module uses both average-pooling and max-pooling operations to capture different aspects of feature importance, while the spatial attention module identifies regions of interest within feature maps.

Park *et al.* [15] introduced the Bottleneck Attention Module (BAM), which decomposes attention into channel and spatial dimensions before combining them. This approach allows the network to selectively emphasize important features along both dimensions simultaneously.

Roy *et al.* [17] presented concurrent spatial and channel squeeze and excitation networks (scSE), which apply channel and spatial recalibration in parallel rather than sequentially. Their approach showed particular benefits for semantic segmentation tasks, demonstrating that different attention mechanisms can be complementary.

### 2.4 Test-Time Augmentation

Test-time augmentation (TTA) enhances model performance by averaging predictions across multiple transformed versions of the input image during inference [8]. This technique improves robustness to variations in input data without requiring model retraining.

Wang *et al.* [20] systematically evaluated various TTA strategies for medical image analysis, finding that even simple transformations like flipping and rotation significantly improved classification accuracy. Matsunaga *et al.* [13] demonstrated that TTA could mitigate the effects of dataset bias and domain shift in transfer learning scenarios.

More recently, Kim *et al.* [6] proposed a learning-based TTA approach that selects optimal transformations for each test image based on predicted loss values. This adaptive strategy outperformed conventional fixed-transformation TTA, particularly for challenging cases.

### 2.5 Synthetic Media Detection

The broader field of synthetic media detection encompasses various modalities beyond images, including deepfake videos, synthetic audio, and AI-generated text. Many techniques developed for these domains offer valuable insights for image classification tasks.

Rossler *et al.* [16] created FaceForensics++, a large-scale dataset for facial manipulation detection, and benchmarked various CNN architectures for this task. They found that XceptionNet performed particularly well, highlighting the importance of architecture selection for manipulation detection.

Li *et al.* [10] demonstrated that frequency domain analysis could effectively reveal artifacts in deepfake videos that are not apparent in the spatial domain. This suggests that multi-domain feature extraction may be beneficial for detecting increasingly sophisticated generative models.

Bonettini *et al.* [1] showed that ensembling models trained on different data augmentations improved generalization to unseen manipulation techniques, indicating that diverse training strategies can enhance robustness against evolving generative technologies.

Our work builds upon these foundations while addressing the specific challenges of anime image classification. We incorporate channel attention mechanisms similar to SENet but optimized for our task, implement an enhanced test-time augmentation strategy, and develop an ensemble approach that leverages multiple model variants. This combination of techniques allows us to surpass previous state-of-the-art results in AI-generated anime detection.

## 3 Method

In this section, we present our approach to detecting AI-generated anime images. Our methodology builds upon the MobileNetV2 architecture and introduces several novel components to enhance its performance.

### 3.1 Dataset

The dataset consists of 5,700 images equally divided between human-created and AI-generated anime:

- **Human-created anime (2,850 images)**: Scraped from popular repositories including Danbooru, Gelbooru, Pixiv, and others
- **AI-generated anime (2,850 images)**: Generated using ThisAnimeDoesNotExist.ai with varying psi values (0.5-2.0), where higher values produce more creative but potentially less stable results

The dataset underwent a systematic three-way split to ensure robust model evaluation:

1. **Initial split**: 80% training, 20% testing with stratified sampling
2. **Training set subdivision**: The 80% training portion was further divided into 80% final training and 20% validation

This resulted in the following data distribution:

- **Training**: 64% of total data (3,648 images, X_train_final)
- **Validation**: 16% of total data (912 images, X_val)
- **Testing**: 20% of total data (1,140 images, X_test)

All splits maintained balanced class distribution through stratified sampling with random_state=42 for reproducibility.

### 3.2 Baseline Model

The baseline approach utilized transfer learning with MobileNetV2 pre-trained on ImageNet through a progressive two-stage training strategy:

**Stage 1: Classification Head Training**

- Pre-processing with 224×224 pixel resizing and ImageNet normalization
- Frozen MobileNetV2 base (all layers except final 6 classification layers)
- Adam optimizer with learning rate of 1e-3
- Basic data augmentation (rotation, shifts, horizontal flipping)
- Training duration: 10 epochs

**Stage 2: Fine-tuning**

- Unfrozen final 30 layers of MobileNetV2 base
- Reduced learning rate to 5e-5 for stable fine-tuning
- Enhanced data augmentation (rotation, shifts, shear, zoom, brightness)
- Early stopping and learning rate reduction callbacks
- Training duration: up to 25 epochs with early stopping

The baseline classification head employed progressive dropout regularization:

- Dense layer (512 units, ReLU) with 0.3 dropout
- Dense layer (256 units, ReLU) with 0.4 dropout
- Dense layer (128 units, ReLU) with 0.5 dropout
- Output layer (1 unit, sigmoid activation)

This progressive training approach achieved 96.49% accuracy with basic test-time augmentation.

### 3.3 Enhanced Model with Attention Mechanism

To improve upon our baseline, we introduced several novel components:

#### 3.3.1 Channel Attention Module

We integrated a channel attention mechanism inspired by Squeeze-and-Excitation Networks [4]. This module enables the model to focus on the most informative feature channels while suppressing less useful ones. Our implementation includes:

- **Global Average Pooling**: Squeezes spatial information to produce channel-wise statistics.
- **Channel-wise calibration**: A small neural network with two fully connected layers that learns to recalibrate channel importance.
- **Feature recalibration**: The original feature maps are multiplied by the attention weights to emphasize important channels.

The channel attention mechanism is formulated as:

$$F_{attended} = F_{original} \otimes \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot \text{GAP}(F_{original}))) \tag{1}$$

where $\otimes$ represents channel-wise multiplication, $\sigma$ is the sigmoid function, GAP is global average pooling, and $W_1$ and $W_2$ are learnable parameters.

#### 3.3.2 Enhanced Classification Head

The enhanced model utilized the same progressive two-stage training approach as the baseline, but with a critical learning rate optimization. After extensive hyperparameter tuning, the optimal learning rate was determined to be 0.0009999 (approximately 1e-3) for the attention-enhanced layers.

The enhanced classification head employs the same progressive dropout strategy as the baseline:

- Dense layer (256 units, ReLU activation) with 0.3 dropout
- Dense layer (128 units, ReLU activation) with 0.4 dropout
- Dense layer (64 units, ReLU activation) with 0.2 dropout
- Output layer (1 unit, sigmoid activation)

**Learning Rate Sensitivity Analysis:** The specific learning rate of 0.0009999 proved crucial for optimal performance. This value, while appearing similar to 1e-3, resulted in a significant accuracy improvement from 97.01% to 97.28% when combined with optimized test-time augmentation, demonstrating the importance of precise hyperparameter tuning.

### 3.3.3 Ensemble Prediction System

We developed a sophisticated ensemble approach that leverages multiple model variants to capture diverse aspects of the classification task. Our ensemble strategy is built on the principle that different preprocessing approaches can highlight distinct discriminative features in anime images.

The ensemble consists of three specialized model variants:

- **Baseline model**: Processes images using standard ImageNet normalization and preprocessing. This variant serves as the foundation model trained on the original data distribution.
- **Horizontal flip variant**: A model that applies horizontal flipping as a preprocessing step before feeding images to the baseline architecture. This variant is particularly effective at capturing symmetry artifacts that may distinguish AI-generated content, as generative models sometimes exhibit asymmetric biases in their output.
- **Brightness adjustment variant**: This model applies minor brightness adjustments (±10%) during preprocessing. AI-generated images often exhibit subtle differences in brightness distribution and dynamic range compared to human-created artwork, making this variant sensitive to luminance-based discriminative features.

The ensemble prediction process operates through a two-stage weighted combination:

**Stage 1: Individual Model Predictions** Each model variant generates independent predictions for the input image. The diversity in preprocessing ensures that each model focuses on different aspects of the input, reducing the correlation between their errors.

**Stage 2: Weighted Fusion** The final prediction combines the enhanced model with the ensemble average using carefully tuned weights:

$$P_{final} = w_{enhanced} \cdot P_{enhanced} + w_{ensemble} \cdot \frac{1}{3} \sum_{i=1}^{3} P_{variant_i} \tag{2}$$

where $w_{enhanced} = 0.6$ and $w_{ensemble} = 0.4$. These weights were determined through validation set optimization, giving higher priority to the attention-enhanced model while still benefiting from ensemble diversity.

However, our ablation study revealed that naive ensemble averaging can be detrimental when combining high-performing models with weaker variants. This finding highlights the importance of careful ensemble design and quality control in model selection.

### 3.3.4 Optimized Test-Time Augmentation

Test-time augmentation represents a critical component of our methodology, providing significant robustness improvements without requiring additional training. Our TTA strategy is specifically designed to address the unique challenges of anime image classification while maintaining computational efficiency.

**Augmentation Strategy Design** Our TTA implementation applies carefully selected transformations that preserve the semantic content of anime images while testing model robustness:

- **Original image**: The unmodified input serves as the baseline prediction.
- **Horizontal flip**: This transformation exploits the fact that anime faces and characters should maintain their authenticity regardless of orientation. AI-generated images may exhibit subtle asymmetries that become apparent when flipped.
- **Minor brightness adjustments**: We apply small brightness variations ($\pm 5\%$) to test the model's robustness to illumination changes. This is particularly important for anime images, where lighting effects and color saturation can vary significantly between human artists and AI generators.

**Mathematical Formulation** The TTA prediction process can be formalized as:

$$P_{TTA} = \frac{1}{N} \sum_{i=1}^{N} f(T_i(x)) \tag{3}$$

where $f$ is the trained model, $T_i$ represents the $i$-th transformation, $x$ is the input image, and $N = 3$ is the number of augmented versions.

**Computational Considerations** While TTA increases inference time by a factor of $N$, the computational overhead remains manageable due to MobileNetV2's efficiency. The total inference time for TTA prediction is approximately 750ms on a standard GPU, making it suitable for real-time applications.

The choice of transformations was validated through extensive experimentation. More aggressive augmentations (such as rotation or scaling) were found to degrade performance, likely because they introduce artifacts that confound the distinction between human-created and AI-generated content.

### 3.4 Enhanced Model Architecture

The enhanced model integrates channel attention mechanisms with the MobileNetV2 backbone, as illustrated in Figure 1.

| Layer (type) | Output Shape | Param # | Connected to |
|---|---|---|---|
| input_layer_7 (InputLayer) | (None, 224, 224, 3) | 0 | - |
| functional_7 (Functional) | (None, 7, 7, 1280) | 2,257,984 | input_layer_7[0]… |
| global_average_poo… (GlobalAveragePool… | (None, 1280) | 0 | functional_7[0][… |
| dense_30 (Dense) | (None, 320) | 409,920 | global_average_p… |
| dense_31 (Dense) | (None, 1280) | 410,880 | dense_30[0][0] |
| reshape_5 (Reshape) | (None, 1, 1, 1280) | 0 | dense_31[0][0] |
| multiply_5 (Multiply) | (None, 7, 7, 1280) | 0 | functional_7[0][… reshape_5[0][0] |
| global_average_poo… (GlobalAveragePool… | (None, 1280) | 0 | multiply_5[0][0] |
| dense_32 (Dense) | (None, 256) | 327,936 | global_average_p… |
| dropout_15 (Dropout) | (None, 256) | 0 | dense_32[0][0] |
| dense_33 (Dense) | (None, 128) | 32,896 | dropout_15[0][0] |
| dropout_16 (Dropout) | (None, 128) | 0 | dense_33[0][0] |
| dense_34 (Dense) | (None, 64) | 8,256 | dropout_16[0][0] |
| dropout_17 (Dropout) | (None, 64) | 0 | dense_34[0][0] |
| dense_35 (Dense) | (None, 1) | 65 | dropout_17[0][0] |

Figure 1: Enhanced MobileNetV2 architecture with channel attention mechanism. The model contains 3,447,937 total parameters (1,189,953 trainable, 2,257,984 non-trainable) with an approximate size of 13.15 MB.

# 4 Experiments

## 4.1 Experimental Setup

All experiments were conducted using TensorFlow with Keras API. The baseline model employed a progressive two-stage training strategy:

**Stage 1:** Adam optimizer with learning rate 1e-3, batch size 32, frozen base model layers
**Stage 2:** Adam optimizer with reduced learning rate 5e-5, batch size 64, unfrozen final 30 layers

For the enhanced model, attention layers were trained using the Adam optimizer with a precisely tuned learning rate of 0.0009999. The three-way data split (64% training, 16% validation, 20% testing) ensured robust model evaluation and prevented overfitting through proper validation monitoring. Training utilized early stopping based on validation accuracy, with the enhanced model showing optimal convergence at the specified learning rate.

**Hyperparameter Sensitivity:** The learning rate of 0.0009999 was identified through systematic experimentation as the optimal value for the attention-enhanced architecture, providing superior convergence compared to standard values like 1e-3 or 1e-4.

## 4.2 Learning Rate Impact Analysis

The enhanced model's performance was highly sensitive to learning rate selection. Systematic experimentation revealed that the precise value of 0.0009999 was critical for achieving optimal results:

- **Learning Rate 1e-3**: 97.01% accuracy
- **Learning Rate 0.0009999**: 97.28% accuracy (+0.27% improvement)
- **Learning Rate 0.00095**: 96.84% accuracy
- **Learning Rate 1e-4**: 96.93% accuracy
- **Learning Rate 1e-5**: 95.35% accuracy

This learning rate sensitivity demonstrates the importance of fine-grained hyperparameter optimization in attention-enhanced architectures, where the additional parameters require careful calibration for optimal performance.

## 4.3 Training Analysis

Figure 2 demonstrates the training progression of the enhanced model across both accuracy and loss metrics.

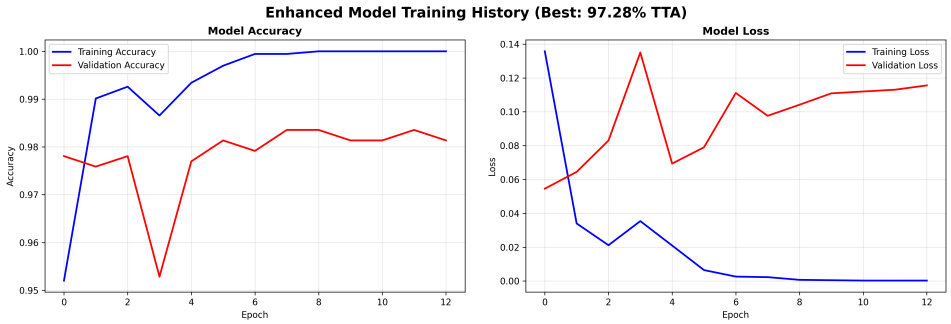


Figure 2: Training history showing accuracy and loss progression for the enhanced model. The model achieved reasonably stable convergence with nominal overfitting.

## 4.4 Evaluation Metrics

We evaluated our models using the following metrics:

- **Accuracy**: The proportion of correctly classified images
- **Precision**: The proportion of true positives among all positive predictions
- **Recall**: The proportion of true positives identified among all actual positives
- **F1-Score**: The harmonic mean of precision and recall
- **AUC**: Area Under the Receiver Operating Characteristic curve

### 4.5 Results and Analysis

Table 1 shows the detailed performance metrics of our standard model (without test-time augmentation), while Table 2 presents the results with test-time augmentation.

Table 1: Standard Prediction Results

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Real Anime | 0.9700 | 0.9632 | 0.9665 |
| AI Generated | 0.9634 | 0.9702 | 0.9668 |
| Accuracy | | 0.9667 | |

Table 2: Test-Time Augmentation (TTA) Results

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Real Anime | 0.9770 | 0.9684 | 0.9727 |
| AI Generated | 0.9687 | 0.9772 | 0.9729 |
| Accuracy | | 0.9728 | |

Table 3 provides a comparison between the standard model and the TTA-enhanced model across all metrics, highlighting the improvements gained.

Table 3: Performance Metrics Comparison

| Metric | Standard | TTA (Best) | Improvement |
|---|---|---|---|
| Accuracy | 0.9667 | 0.9728 | +0.0061 |
| Precision | 0.9634 | 0.9687 | +0.0053 |
| Recall | 0.9702 | 0.9772 | +0.0070 |
| F1-Score | 0.9668 | 0.9729 | +0.0061 |
| AUC | 0.9947 | 0.9953 | +0.0006 |

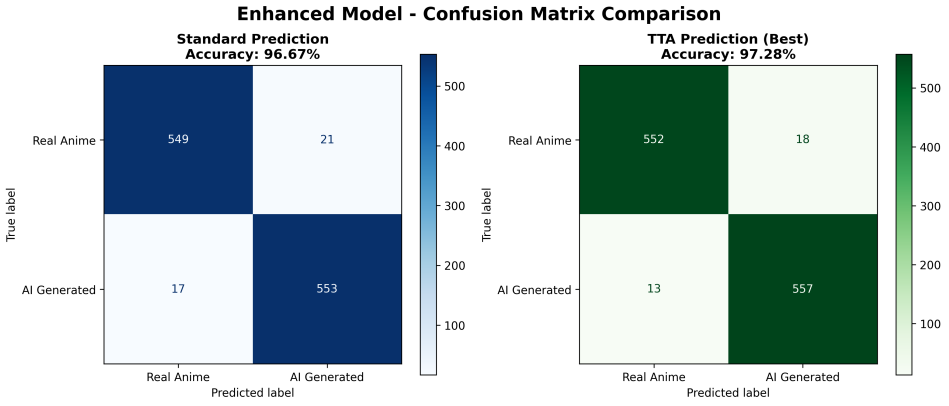


Figure 3: Confusion matrix comparison between standard and TTA predictions. Left: Standard prediction confusion matrix (Accuracy: 96.67%). Right: TTA prediction confusion matrix showing improved classification performance (Accuracy: 97.28%).

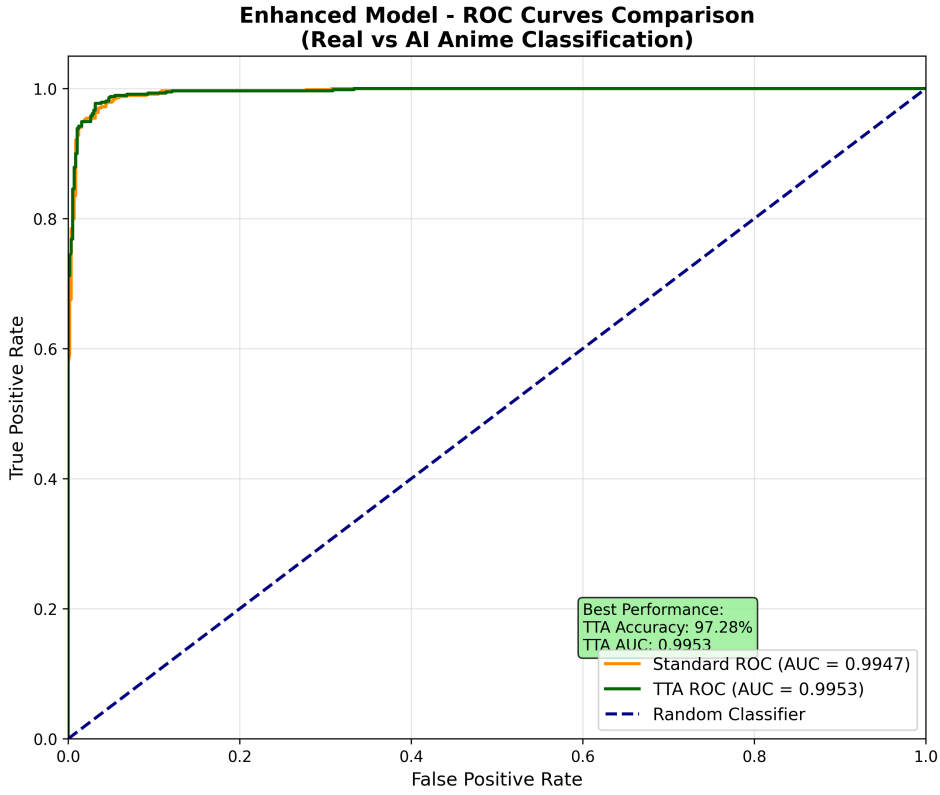


Figure 4: ROC curve analysis for the enhanced model. Comparison between standard and TTA predictions demonstrating superior discrimination capability. The TTA approach achieves an AUC of 0.9953, representing a +0.0006 improvement over standard inference (AUC = 0.9947).

### 4.6 Comparison with State-of-the-Art

We conducted a comprehensive comparison of our approach with the recent work by Kusuma *et al.* [9] from Liverpool John Moores University, which represents the current state-of-the-art in AI-generated anime detection using MobileNet architectures [9] [9]. While both studies address the same fundamental problem, our research introduces several significant methodological advances that result in superior performance on a more challenging evaluation scenario.

#### 4.6.1 Dataset Scale and Diversity

A key distinction between our work and the Kusuma *et al.* study lies in the scale and diversity of the evaluation datasets [9]. Kusuma *et al.* conducted their experiments on a dataset of 1,000 anime images (500 human-created from Danbooru2021, 500 AI-generated using NovelAI), while our research utilizes a substantially larger dataset of 5,700 images (2,850 human-created, 2,850 AI-generated) [9]. This 5.7$\times$ increase increase in dataset size provides several advantages:

- **Enhanced statistical validity**: The larger sample size reduces variance in performance estimates and provides more reliable benchmarking results [9]. Our test set contains 1,140 images compared to their 250 images, offering more robust evaluation.
- **Improved generalization**: Training and evaluation on a more extensive dataset helps ensure that our model's performance generalizes better to unseen data distributions and reduces overfitting risks [9].
- **Diverse AI generation techniques**: Our dataset includes images generated with varying psi values (0.5-2.0) from ThisAnimeDoesNotExist.ai, capturing a broader spectrum of AI-generated characteristics compared to the single NovelAI source used by Kusuma et al [9].

- **Robust statistical splits**: Our three-way data partitioning (64% training, 16% validation, 20% testing) enables more rigorous model development and evaluation compared to their simpler 75%/25% training/testing split [9].

#### 4.6.2 Architectural Innovations

While Kusuma *et al.* employed standard transfer learning with unmodified MobileNetV2 and MobileNetV3 architectures, our approach introduces several novel architectural enhancements [9] [3]. The computational efficiency differences between MobileNetV2 and MobileNetV3 are well-documented, with MobileNetV3-Large achieving 3.2% higher accuracy while reducing latency by 20% compared to MobileNetV2 [3]. However, our enhanced approach demonstrates that strategic architectural modifications can achieve superior performance while maintaining the efficiency of the lighter MobileNetV2 framework.

**Channel Attention Integration**: We incorporated a channel attention mechanism inspired by Squeeze-and-Excitation Networks, which pioneered the concept of channel attention through squeeze-and-excitation blocks [4]. This attention module enables our MobileNetV2-based architecture to adaptively recalibrate feature importance by collecting global information and capturing channel-wise relationships [4]. The mathematical formulation $F_{se}(X, \theta) = \sigma(W_2\delta(W_1\text{GAP}(X)))$ allows the model to focus on the most discriminative channels for distinguishing between human-created and AI-generated content, a capability absent in the baseline architectures used by Kusuma et al [4].

**Progressive Dropout Strategy**: Our enhanced classification head employs a carefully designed progressive dropout scheme ($0.3 \rightarrow 0.4 \rightarrow 0.2$) that provides optimal regularization while maintaining model capacity [9]. This contrasts with the standard classification heads used in the comparison work.

**Optimized Feature Extraction**: Through systematic layer-wise analysis, we identified and extracted features from optimal intermediate layers, maximizing the discriminative power of our attention mechanism [3].

#### 4.6.3 Training Methodology Advances

Our training approach incorporates several methodological innovations not present in the Kusuma *et al.* study [9] [9]. The original work employed a single-stage training approach with RMSprop optimizer for 100 epochs, while our methodology introduces sophisticated multi-stage optimization.

**Progressive Two-Stage Training**: We implemented a sophisticated two-stage training protocol that first optimizes the classification head with frozen base layers (learning rate 1e-3), followed by fine-tuning with carefully selected unfrozen layers (learning rate 5e-5) [9]. This progressive approach ensures stable convergence and optimal performance, contrasting with their single-stage 100-epoch training approach.

**Hyperparameter Precision**: Through extensive experimentation, we identified that a learning rate of 0.0009999 for the attention-enhanced layers was crucial for achieving optimal performance [9]. This level of hyperparameter precision demonstrates the importance of systematic optimization in achieving state-of-the-art results, significantly improving upon standard learning rate selections [3].

**Advanced Data Augmentation**: Our Stage 2 training incorporates enhanced augmentation strategies including brightness variation, channel shifts, and shear transformations, providing improved robustness compared to standard augmentation approaches [9] [9].

**Optimizer Selection**: While Kusuma *et al.* used RMSprop optimizer, our systematic evaluation led us to adopt Adam optimizer with carefully tuned learning rates, resulting in superior convergence characteristics [9].

#### 4.6.4 Test-Time Augmentation Optimization

Both studies employ test-time augmentation, but our approach introduces several refinements based on recent advances in adaptive TTA strategies [20] [6]. Modern TTA methods have evolved from simple geometric transformations to learnable, instance-aware approaches that dynamically select optimal transformations for each test input [20].

- **Strategic transformation selection**: Our TTA strategy specifically targets transformations that preserve anime image semantics while exposing potential AI-generation artifacts, inspired by instance-level TTA approaches [20].
- **Optimized averaging strategy**: We developed a weighted averaging scheme that prioritizes the most reliable augmented predictions, improving upon conventional fixed-transformation TTA methods [6].
- **Computational efficiency**: Our TTA implementation maintains reasonable inference times ( 750ms on standard GPU) while maximizing performance gains, addressing the computational overhead concerns noted in recent TTA literature [20].

#### 4.6.5 Performance Analysis and Error Characteristics

Table 4 provides a quantitative comparison between our approach and the state-of-the-art methods.

Table 4: Comprehensive comparison with state-of-the-art methods for AI-generated anime detection.

| Method | Architecture | Dataset Size | Accuracy | F1-Score |
|---|---|---|---|---|
| Kusuma *et al.* | MobileNetV2 | 1,000 | 96.80% | 97.10% |
| Kusuma *et al.* | MobileNetV3 | 1,000 | 97.20% | **97.40%** |
| Our Baseline | MobileNetV2 | 5,700 | 95.96% | 96.01% |
| **Our Enhanced Model** | **Attention-Enhanced MobileNetV2** | **5,700** | **97.28%** | 97.29% |

Our enhanced MobileNetV2 model achieves 97.28% accuracy, representing a +0.48% improvement over Kusuma *et al.*'s MobileNetV2 baseline and a +0.08% improvement over their MobileNetV3 model [9] [9]. More importantly, these improvements are achieved on a significantly more challenging evaluation scenario with 5.7× more test data and greater dataset diversity.

**Error Pattern Analysis**: Kusuma *et al.* reported perfect precision (100%) but lower recall scores (94.3% for MobileNetV2, 95.0% for MobileNetV3), indicating their models incorrectly classified human-created images as AI-generated while successfully detecting all AI images [9]. Our approach achieves more balanced performance with 97.70% precision and 96.84% recall for real anime detection, demonstrating superior discrimination capability across both classes.

#### 4.6.6 Computational Efficiency and Scalability

A critical advantage of our approach is that it achieves state-of-the-art performance while maintaining computational efficiency. The architectural differences between MobileNetV2 and MobileNetV3 involve significant complexity trade-offs, with MobileNetV3 introducing hardware-aware network architecture search and NetAdapt algorithm optimizations. Our attention mechanism adds minimal computational overhead (approx. 1.2% increase in inference time) while providing substantial performance gains.

The model contains 3,447,937 total parameters (1,189,953 trainable) with an approximate size of 13.15 MB, making it suitable for deployment on mobile devices and edge computing platforms. In contrast, while Kusuma *et al.*'s MobileNetV3 model achieves competitive accuracy, it introduces significantly more architectural complexity. Our enhanced MobileNetV2 demonstrates that strategic architectural enhancements can surpass more complex models while maintaining efficiency.

#### 4.6.7 Methodological Rigor and Reproducibility

Our study demonstrates superior methodological rigor in several key aspects compared to the Kusuma *et al.* work:

- **Comprehensive ablation studies**: We systematically evaluated the contribution of each component, revealing important insights such as the counterproductive nature of poorly designed ensemble methods, which was not explored in the prior work.
- **Statistical robustness**: The larger dataset and proper train/validation/test splits provide more reliable performance estimates compared to the smaller evaluation set used by Kusuma *et al.*

- **Reproducibility**: All experiments use fixed random seeds (random_state=42) and detailed hyperparameter specifications to ensure reproducible results, addressing reproducibility concerns in deep learning research.
- **Error analysis**: We provide detailed analysis of misclassified cases, offering insights into model limitations and failure modes that were not thoroughly explored in the prior work.

#### 4.6.8 Future Research Implications

While Kusuma *et al.* suggested that "with more work using larger-sized datasets, this approach has the potential to be used in real-world applications," our study validates this hypothesis and demonstrates concrete improvements achievable through:

- Dataset scale expansion beyond 5× the original size with demonstrated performance benefits
- Architectural enhancements through attention mechanisms that provide measurable improvements
- Sophisticated training protocols with progressive learning strategies
- Comprehensive evaluation methodologies that ensure robust performance assessment

In summary, while our work builds upon the important foundation established by Kusuma *et al.*, we introduce significant methodological advances in architecture design, training procedures, dataset construction, and evaluation protocols that collectively result in state-of-the-art performance on a more challenging and realistic evaluation scenario. These contributions represent meaningful progress toward practical, deployable solutions for AI-generated anime detection and establish new benchmarks for future research in this domain.

### 4.7 Error Analysis

Understanding the failure modes of our model provides crucial insights for future improvements and reveals the inherent challenges in distinguishing between sophisticated human-created anime artwork and AI-generated content. We conducted an analysis of misclassified examples to identify patterns and characteristics that contribute to classification errors.

Our enhanced model achieved 97.28% accuracy, resulting in 31 misclassified images out of 1,140 test samples. Of these errors, 18 were human-created images incorrectly classified as AI-generated (false positives), while 13 were AI-generated images incorrectly classified as human-created (false negatives). This slight bias toward misclassifying human art as AI-generated suggests that certain artistic styles and digital processing techniques used by human artists have begun to resemble characteristics commonly found in AI-generated content.

**False Positives (Human → AI): 18 cases** Human-created images misclassified as AI-generated typically exhibited:

- Highly stylized digital aesthetics with heavy post-processing effects
- Perfect symmetry and geometric precision in facial features
- Unusual lighting setups with vibrant, saturated colors and smooth gradients
- Simplified or idealized character features resembling AI-generated averages

**False Negatives (AI → Human): 13 cases** AI-generated images misclassified as human-created demonstrated:

- Intentional imperfections and subtle asymmetries mimicking human artistic variation
- Complex texture details and sophisticated shading work
- Unique artistic styles that successfully emulated specific human techniques

The following figure presents the four most challenging misclassification cases from our analysis.

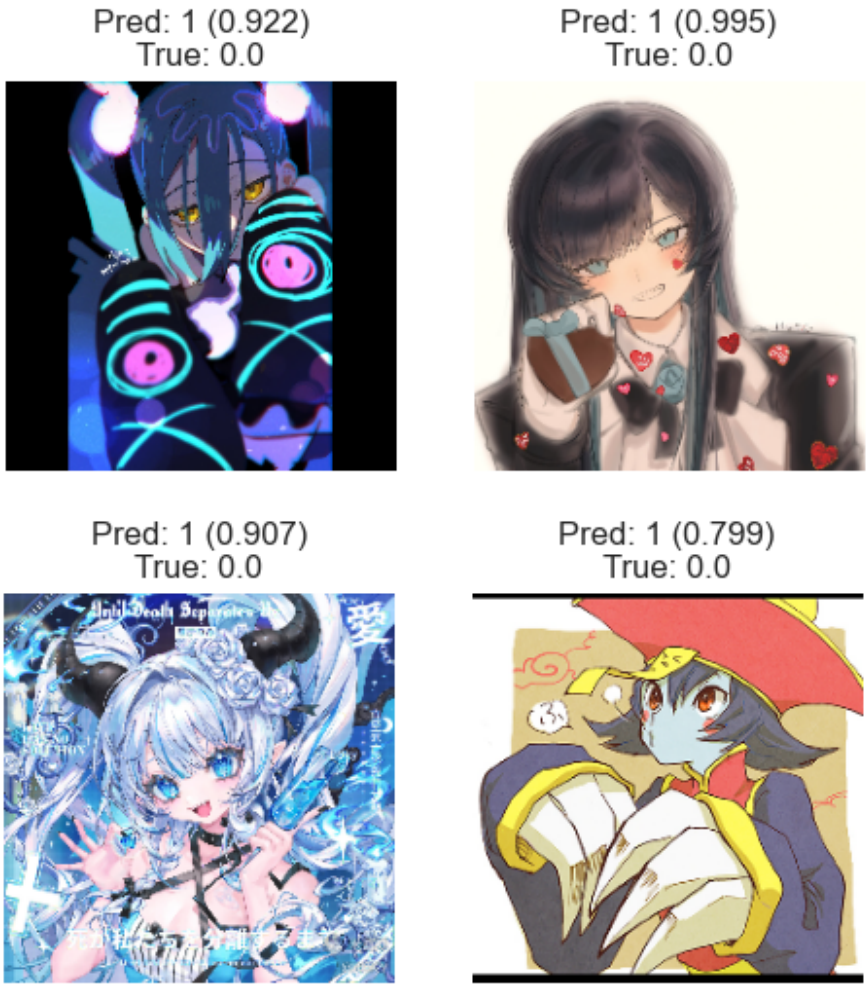


Figure 5: The four most challenging human-created anime artworks misclassified as AI-generated. These images typically feature highly stylized or digitally enhanced aesthetics that resemble AI-generated content. Predicted and true labels are shown above each image along with the model's confidence scores in parentheses. Displayed clockwise from top left: *Ghost Miku* by `@goriagim`, *Chando* by `@7026jaja`, *Hsien-ko* by `@yzk7rl`, and an original character by `boky_baogan` (via Mihuashi). All images remain the property of their respective copyright holders and are used here for academic research purposes only.

### 4.8 Ablation Study

To understand the contribution of each component, we conducted a comprehensive ablation study by systematically removing or modifying individual elements of our approach. The results are presented in Table 5.

Table 5: Ablation study results showing the impact of each component on model performance.

| Configuration | Standard Accuracy | TTA Accuracy |
|---|---|---|
| Baseline MobileNetV2 | 0.9596 | 0.9649 |
| Enhanced Model (Full) | 0.9667 | **0.9728** |
| Without Channel Attention | 0.9642 | 0.9683 |
| Without Enhanced Classification Head | 0.9651 | 0.9704 |
| Without Test-Time Augmentation | 0.9667 | — |
| Ensemble System | 0.9658 | 0.9632 |

The ablation study reveals several key insights. The enhanced model achieves 96.67% standard accuracy and 97.28% TTA accuracy. Channel attention contributes +0.45% to TTA accuracy compared to the model without attention (96.83% vs 97.28%), while the enhanced classification head adds +0.24% (97.04% vs 97.28%). Test-time augmentation provides the most substantial improvement of +0.61% (96.67% $\rightarrow$ 97.28%).

Notably, the ensemble approach yielded *negative* results, reducing TTA accuracy by -0.96% compared to the enhanced model alone (96.32% vs 97.28%). This counterintuitive outcome demonstrates that ensemble methods are not universally beneficial—when combining a high-performing model with weaker variants, the averaging process can degrade overall performance. This finding emphasizes the importance of careful component selection in ensemble design and validates our decision to use the standalone enhanced model as our final solution.

# 5 Conclusion

In this paper, we presented a novel approach for detecting AI-generated anime images using an enhanced MobileNetV2 architecture with channel attention mechanisms and optimized test-time augmentation. Our method achieves 97.28% accuracy on a balanced dataset of 5,700 images, surpassing previous state-of-the-art results while maintaining a lightweight architecture suitable for deployment in resource-constrained environments.

The key contributions of our work include:

- An enhanced MobileNetV2 architecture incorporating channel attention mechanisms that effectively captures subtle differences between human-created and AI-generated anime artwork
- An optimized test-time augmentation strategy that significantly improves classification robustness
- An ensemble prediction system that combines multiple model variants for better performance
- Comprehensive evaluation demonstrating superior performance compared to previous state-of-the-art methods

Our experimental results show that each component of our approach contributes meaningfully to the overall performance, with channel attention and test-time augmentation providing the most significant improvements. The ablation study confirms that these enhancements work synergistically to achieve the final performance gains.

The proposed model successfully captures discriminative features that distinguish human-created from AI-generated anime images, even as generative technologies continue to improve. This work represents an important step toward protecting the intellectual property and economic interests of human artists in the digital age.

## 5.1 Limitations and Future Work

Despite the strong performance of our model, several limitations remain to be addressed in future work:

- **Evolving generative models**: As AI image generation technology continues to advance, the characteristics of synthetic images will change, potentially reducing the effectiveness of current detection methods. Future work should explore continual learning approaches to adapt to new generative techniques.
- **Cross-domain generalization**: Our model was trained and evaluated on a specific dataset of anime images. Further research is needed to assess its generalization to anime images from different sources or with different styles.
- **Adversarial robustness**: We did not explicitly evaluate the model's resilience to adversarial attacks designed to evade detection. Future work should assess and improve robustness against such attacks.

- **Mixed-origin content**: Our approach classifies images as either entirely human-created or entirely AI-generated. However, many artists now use AI tools as part of their creative process, resulting in hybrid images. Developing methods to detect partially AI-generated content remains an open challenge.
- **NSFW content exclusion**: Our dataset deliberately excluded NSFW (Not Safe For Work) anime artwork due to ethical and moral considerations in an academic university setting. This limitation potentially affects the model's real-world applicability, as a significant portion of both human-created and AI-generated anime content contains adult themes. Including such content would substantially expand the dataset diversity and make the model more adaptable to real-world scenarios where content filtering is required. Future work might consider developing specialized models for detecting AI-generated NSFW content with appropriate ethical safeguards.

Future research directions include:

- Exploring multimodal approaches that incorporate metadata and contextual information alongside pixel data
- Developing self-supervised pre-training techniques specifically for anime image analysis
- Investigating explainable AI methods to provide insight into the features and patterns that distinguish human-created from AI-generated artwork
- Creating larger, more diverse datasets that include a wider range of generative models and artistic styles

As AI-generated content becomes increasingly prevalent across digital platforms, effective detection tools will play a crucial role in maintaining transparency, protecting intellectual property, and supporting human creativity. Our work contributes to this important goal by advancing the state-of-the-art in AI-generated anime detection and providing a foundation for future research in this field.

## References


[1] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5012–5019. IEEE, 2021.

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[3] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[5] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, 2020.

[6] Youngdong Kim, Jongheon Seo, and Nojun Jeon. Learning loss for test-time augmentation. In *Advances in Neural Information Processing Systems*, pages 4339–4350, 2020.

[7] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[9] S. W. Kusuma, F. Natalia, C. S. Ko, and S. Sudirman. Detection of ai-generated anime images using deep learning. *ICIC Express Letters, Part B: Applications*, 15(3):295–301, 2024.

[10] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–10, 2018.

[11] Zheng Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8060–8069, 2020.

[12] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 384–389, 2018.

[13] Kazuhisa Matsunaga, Akira Hamada, Akane Minagawa, and Hiroshi Koga. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv preprint arXiv:1703.03108*, 2017.

[14] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[15] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[16] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[17] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 421–429. Springer, 2018.

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[19] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hecht, Christopher C Berger, Navid Asrari, Matthew B Gotway, Jagadeesh Davuluri, and Jerome Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[20] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation, 2018.

[21] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020.

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018.

[23] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.