
MIS548

# SPOT THE FAKE

A comparative analysis of deep learning models for detecting AI generated vs. human images: business cases in accuracy, cost, and latency

MOHAMMED SAMEER SYED
ABHIMANYU PANDEY

MAY 2025

# TABLE OF CONTENTS

# Business Problem Summary

**With AI advancements over the past five years, industries face the critical challenge of distinguishing AI-generated images from authentic human content. As with many deep learning applications, success hinges on thoroughly understanding the specific business problem and available data before selecting an appropriate model. One of the most pressing problems across the industry is that businesses undervalue the importance of comprehending requirements when developing tailored solutions. For example, some scenarios demand accuracy, others require cost efficiency at scale, whilst some prioritise real time responsiveness.**

Selecting an inappropriate model architecture can cause significant immediate risks, but also incurs costs later on. Misclassification errors that propagate incorrect data into the pipeline, degrading the quality of retraining datasets and subsequent models. This data pollution leads to skewed analytics, reduced reliability of predictive models, and ultimately undermines decision-making confidence. The long-term implications for the broader data ecosystem are severe; once bias or inaccuracies enter the pipeline, they become progressively challenging and costly to rectify, undermining competitive advantage and company share price.

Furthermore, choosing the incorrect model can lead to substantial opportunity costs. Failing to capture accurate, low-latency, or cost-effective data compromises the business's ability to adapt rapidly to market changes, eroding future competitive positioning and negatively impacting potential revenue streams. Conversely, selecting the appropriate model architecture significantly enhances data integrity, bolsters the reliability of analytics, and supports strategic decisions, contributing significant value to company performance and shareholder confidence.

To bring to life this objective, we look at three different business problems to which this same exercise of AI/no AI classification could be applied to. Each scenario is selected because different architectures will shine when applied to different business problems, We will demonstrate this by assessing how latency, inference cost efficiency and accuracy play off in each scenario. Although we focus on the application of a ‘spot the fake’ model (AI/not AI) the business problem we approach is one of the most pressing in the industry, and will continue to be of relevance in the coming years.

# Business Problem Summary

## Accuracy: Passport Verification for Visa Applications

The Dinia Passport Office needs a deep learning system to detect fake passport photos in visa applications, ideally requiring around 99% accuracy to prevent identity fraud, human trafficking, or terrorism. False positives delay or deny legitimate applicants, impacting businesses and tourists. Latency and cost are secondary; accuracy is critical.
A model safcrificing accuracy exposes the institution to unacceptable risks in numerous areas.

**Security**: Missed fraud cases increase identity fraud (global cost >US$50B by 2025, Experian, 2023) and security vulnerabilities

**Compliance**: Misclassifications lead to AML/KYC fines (~US$5B globally)

**Operational Costs:** False positives trigger manual reviews (~US$2–5 each)

**Data Integrity**: Errors pollute immigration databases, degrading trust and requiring costly corrections.

**Reputation**: Delays harm business and tourist visa applicants, impacting trust.

A high-accuracy model ensures security, compliance, and trust, minimizing risks and supporting efficient visa processing.

## Cost: Social Media Content Moderation

KlikKlock, a social media video platform, manages a million user-generated video uploads monthly and must tag AI generated content to prevent scams and harmful deepfakes. Neither accuracy nor latency is critical; videos are processed asynchronously, and lower accuracy only moderately impacts user engagement. Given the sheer scale and the fact that this service is provided for free, inference cost efficiency becomes paramount.

If KlikKlock adopts a computationally expensive model, inference costs may become financially unsustainable at scale, risking operational viability. However, choosing a model solely on low cost with minimal accuracy can significantly degrade user engagement. TikTok, for example, generated US$23.3 billion in global ad revenue in 2024, equating roughly to US$230 million in revenue per 1% drop in user engagement or ad impressions.

On the cost side, numerous aspects have significant impacts for the business.
High inference costs strain budgets, potentially exceeding operational viability at scale, diverting funds from ad optimization and reducing ROI. They also hinder platform growth, limiting ability to handle increasing uploads, which constrains ad inventory and revenue potential. Finally, the ability to deliver well priced ads the strong ROI in comparison to other platforms may be affected if costs of running the platform at scale are too high.

## Latency: Real Time Live Streaming Moderation

Live streams are interactive and unedited, which makes them a potential vector for real-time deepfake misuse. Selecting an AI model for livestream content moderation hinges on ultra-low latency to prevent harm, ensure compliance, and maintain business viability. High-latency moderation risks user trust erosion, as seen in Twitch's 2021 #ADayOffTwitch boycott over delayed hate raid responses, and can lead to user flight, traumatizing viewers and driving creators away. Regulatory penalties, like the EU's 2022 Terrorist Content Regulation mandating one-hour takedowns with fines up to 4% of global revenue, make slow responses financially crippling. The moral imperitives of getting latency right are evident in the tragic Christchurch Mosque shooting's 17-minute livestream of a terrorist event, which caused reputational damage to the platform, global regulatory scrutiny and vastly negative impacts on social cohesion and stability. Slow moderation also threatens business continuity, as Parler's 2021 app store removal showed, risking audience loss and advertiser pullbacks. In the data pipeline, low-latency models demand optimized inference and high-quality, real-time data to balance speed and accuracy, while poor data quality risks mis-flagging, further eroding trust. A high-performing model preserves user safety, ecosystem stability, and company valuation by mitigating legal, PR, and revenue risks, ensuring platforms remain trusted and operational.

## Summary of Business Problems

These cases illustrate that model selection must be driven by clear business priorities. Selecting inappropriately balanced models risks immediate operational issues and compounds into long-term data pollution, undermining analytics confidence and strategic decision-making capabilities. The right model architecture for each specific business context significantly enhances data integrity, operational reliability, and ultimately contributes to sustainable competitive advantage.

# Data Overview

**This section overviews the deep learning models assessed in this exercise, detailing their architectures and suitability for AI versus human-generated image classification tasks. We explore the design and hyperparameters of the models.**

The dataset used in this project is specifically curated for distinguishing between AI-generated and authentic human-produced images. This publically available dataset was sourced from Kaggle. the dataset contains diverse images explicitly labelling as either "AI generated (fake)" produced by generative models (usually GANs), or "human-generated (real)," representing genuine photographs.

Each image in the dataset comes with clear labelling : a file name pointing to the exact image location and a binary label indicating its category (0 for AI generated images and 1 for authentic images). Each images was resized to 64x64 pixels to ease computational effort and make the exercise a little more consistent.

The dataset was partitioned into three subsets: a training set to teach the models, a validation set to fine-tune hyperparameters and prevent overfitting, and a test set to evaluate performance objectively. This was done at the level of each model evaluation in alignment with the model at hand.

Pre-processing steps were consistently applied across all models. Images underwent resizing to uniform dimensions and normalization, scaling pixel values to the range [-1, +1]. These steps are standard best practices for image-based neural networks, facilitating faster convergence and stable training across all architectures tested.

Additionally, the dataset's relatively modest size (5,000 images per model training session) could affect generalizability. Larger and more diverse datasets typically yield more robust models capable of accurately handling varied real-world conditions, reducing risks associated with biases inherent in smaller, curated datasets. This is especially relevant in the case where accuracy is essential in the business problem.

Ultimately, the dataset is moderately well designed for initial model training and evaluation, serving as a relatively robust foundation. However, for deployment in high risk scenarios like government ID verification, it would require further enrichment. This could be in the form of higher resolution, expanded size, and reduced bias.

# Project Goals

The overarching goal of this project is to demonstrate the critical importance of selecting the right model architecture based on the specific business problem and dataset at hand. By evaluating three distinct scenarios passport photo verification (accuracy-critical), Klik Klok video moderation (cost efficiency and scale paramount), and livestream content moderation (where latency is most important), this project highlights how mismatched model choices can lead to cascading operational, financial, and reputational costs. The aim is not just to optimise performance in each case, but to show how model selection directly impacts security, user trust, regulatory compliance, and long-term competitive advantage across different data and deployment contexts.

# Deep Learning Overview

**This section overviews the deep learning models assessed in this exercise, detailing their architectures and suitability for AI versus human-generated image classification tasks. We explore the design and hyperparameters of the models.**

## ProGAN Discriminator

The ProGAN discriminator employs a progressive growing GAN (Generative Adversarial Network) architecture, originally designed for high-resolution image synthesis tasks. It incrementally increases layer complexity, effectively balancing computational cost, latency, and accuracy. The model processes input images of size 64x64 through convolutional layers, starting from 32 filters and doubling at each subsequent layer up to 128 filters. Each convolutional layer is followed by a LeakyReLU activation and AveragePooling. Training utilised the Adam optimiser with a learning rate of 2e-4 and beta_1 of 0.5, and a batch size of 128, suitable for efficient real-time, high-quality image classification tasks. The dataset was split using a 60/20/20 ratio (train/validation/test), ensuring balanced evaluation metrics across a representative dataset. Training continued for 50 epochs, a sufficient duration to achieve stable convergence without overfitting.

## StyleGAN Discriminator

StyleGAN's discriminator is based on a convolutional neural network architecture initially created for generating and identifying stylistically consistent, high-fidelity images. It utilises sequential convolutional blocks, each containing two convolutional layers, Pixel Normalisation, and LeakyReLU activations. Spatial dimensions are progressively reduced while feature depth increases from 64 to 512 filters. Key hyperparameters include the Adam optimiser with a learning rate of 2e-4, beta_1 of 0.5, and a batch size of 64. This architecture effectively balances visual accuracy with moderate inference latency. The dataset used a similar 60/20/20 split (train/validation/test) to maintain consistency and fair comparison. The training lasted 50 epochs, balancing the model's complexity with manageable computational overhead.

## GJOB-GAN Discriminator

The GJOB-GAN discriminator features a robust convolutional GAN architecture optimised for reliable binary classification tasks requiring a balanced trade-off between precision and recall. It incorporates multiple convolutional layers with increasing filter counts from 32 to

256, batch normalisation, and a dropout rate of 0.25 for regularisation. Dense layers are included, reaching up to 512 neurons, enhancing discriminative capability. Training hyperparameters included the Adam optimiser (learning rate 2e-4, beta_1 0.5), batch size of 32, and epochs set to 50, making it suitable for applications demanding high accuracy and reliability at manageable computational costs. The data split here (60/20/20) mirrored that of other GAN architectures to ensure comparable benchmarking and generalisation ability.
proposal is aimed at attracting potential clients with what a company sells. It's a document in either digital or printed form that explains product or service features,

## Timm-based CNN Model

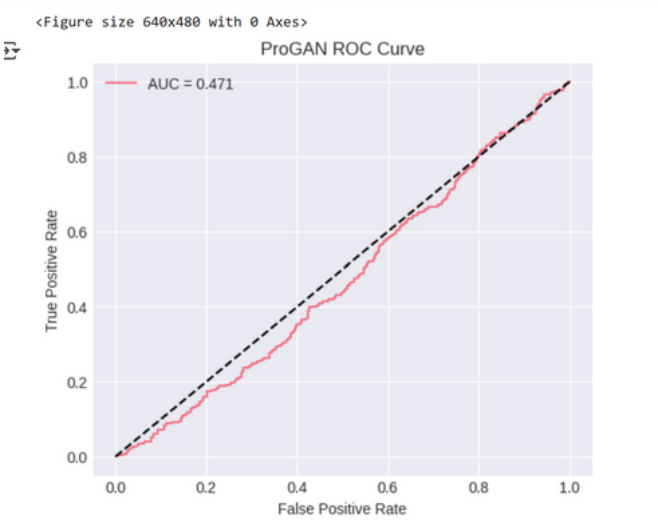
The assignment utilised a CNN model based on the PyTorch Image Models (timm) library, specifically leveraging architectures such as EfficientNet and ResNet. EfficientNet was designed for efficient computation through optimal scaling of network depth, width, and resolution, effectively targeting scenarios demanding low latency and high accuracy. ResNet introduced residual connections to mitigate performance degradation in deeper networks, beneficial for complex visual tasks. Key hyperparameters involved fine-tuning pre-trained weights with a batch size consistent with best practice guidelines for timm-based models, achieving computational efficiency suitable for general business applications. This model used a stratified split of 60/20/20 to ensure diverse and representative training, validation, and testing sets. The number of epochs was chosen to balance fine-tuning effectiveness with computational practicality, typically fewer than the GAN models due to leveraging pre-trained weights.

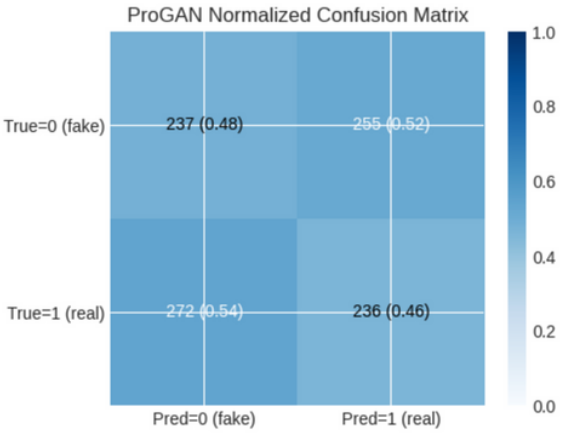
Each model's architecture and hyperparameters were strategically selected to align with specific business scenarios, emphasising optimal accuracy, minimal latency, and cost-effective inference to meet immediate technical requirements and support broader strategic objectives.

# Results

## ProGAN Results Summary

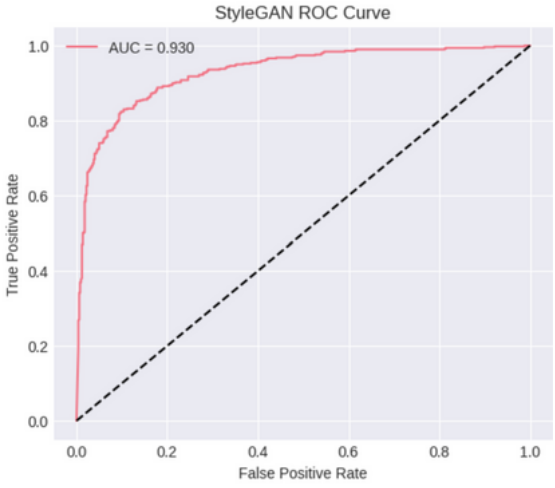
ProGAN delivered relatively poor performance, with an accuracy of only 47% and an AUC of 0.471, indicating the model performs approximately at random guessing level for this specific classification task. Reviewing the training and validation loss curves, the loss decreased steadily over epochs, but plateaued around epoch 15, suggesting limited learning beyond this point. The confusion matrix highlighted evenly poor recognition across both fake and real classes, reflecting a fundamental weakness in distinguishing subtle features. Given ProGAN's progressive GAN architecture, initially designed to handle complex image synthesis tasks with incremental layer complexity, it appears the model's inherent architectural strengths did not translate effectively to binary classification, despite using a well-tuned Adam optimiser (learning rate 2e-4, batch size 128).



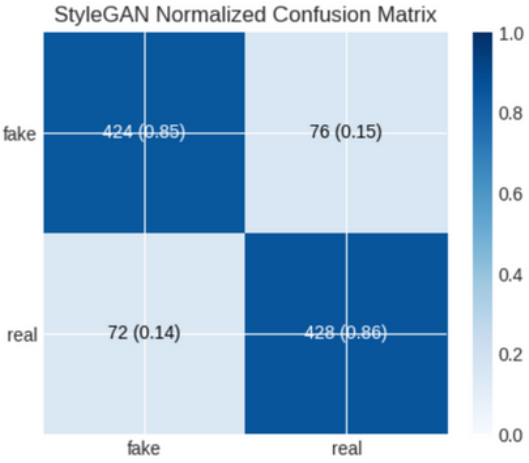

## StyleGAN Results Summary

StyleGAN demonstrated strong performance with an accuracy of 85% and an impressive AUC of 0.930, indicating excellent model discrimination capabilities. Training loss decreased dramatically, approaching near-zero levels, suggesting that the model effectively learned detailed image features during training.



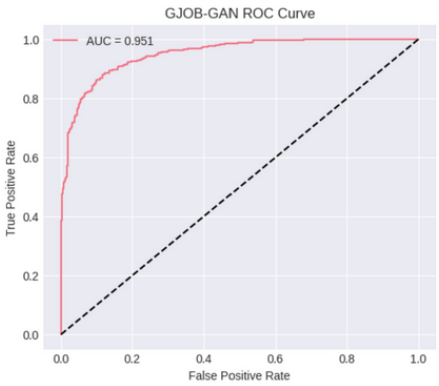
However, the validation loss showed signs of divergence after approximately 20 epochs, highlighting potential overfitting and suggesting that fewer epochs or stronger regularisation might have improved generalisation further.

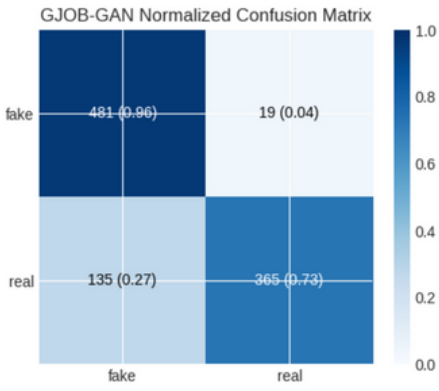
The confusion matrix indicates robust detection capability, correctly classifying around 85% of both real and fake images. StyleGAN's superior results, driven by its structured convolutional blocks and pixel normalisation, clearly outperform ProGAN's near-random results, which struggled significantly with distinguishing between classes.




## GJOB-GAN Results Summary

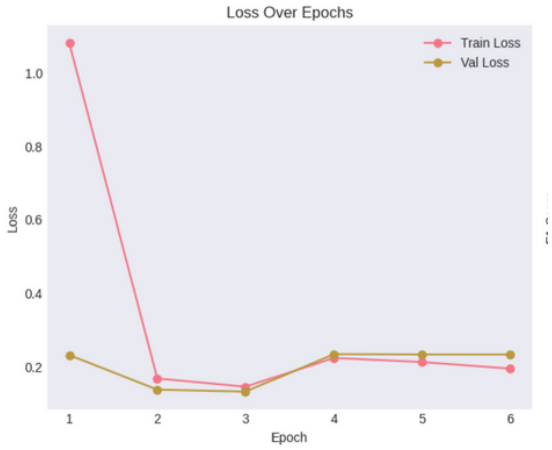
GJOB-GAN exhibited robust performance with an accuracy of 85% and a superior AUC of 0.951, highlighting its strong classification ability. The training loss steadily decreased, indicating effective model learning; however, significant fluctuations and consistently high validation loss throughout training suggest notable overfitting or instability. The confusion matrix showed GJOB-GAN had exceptional precision and recall for detecting fake images (96% recall), although it performed moderately on real images (73% recall). This asymmetry in performance reflects GJOB-GAN's architecture—particularly its dense layers, batch normalisation, and dropout (0.25)—which enabled good discrimination yet potentially caused biased detection towards the fake class. Compared to StyleGAN, GJOB-GAN achieved slightly better overall AUC but showed more significant overfitting concerns.

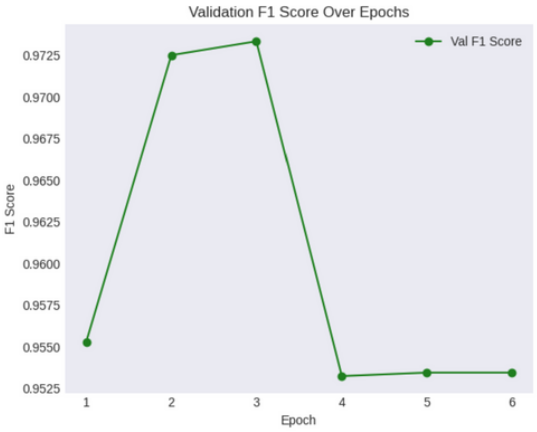


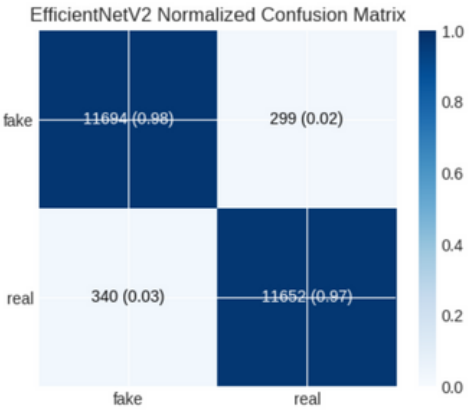



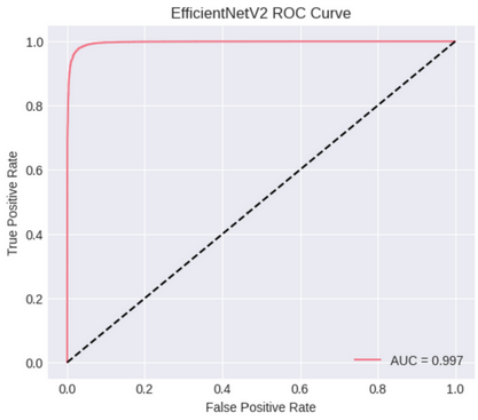
# EfficientNetV2 (Timm) Results Summary

EfficientNetV2 demonstrated exceptional performance, achieving an accuracy of 97%, an F1 score of 0.973, and an outstanding AUC of 0.997, clearly surpassing the previously analysed models. The training and validation loss converged swiftly and stabilised after only a few epochs, highlighting EfficientNetV2's effective utilisation of pre-trained weights, minimal overfitting, and excellent generalisation. The confusion matrix revealed near-perfect classification, correctly identifying 98% of fake images and 97% of real images, reflecting well-balanced and robust model capabilities. This remarkable performance reflects the architecture's inherent strengths particularly the optimal scaling of model depth, width, and resolution in EfficientNet, and efficient training through fine-tuning—resulting in superior accuracy and generalisation compared to the GAN-based models (ProGAN, StyleGAN, and GJOB-GAN).



# Results in Business Context

## Accuracy Scenario: Passport Verification

In the passport verification scenario, accuracy is paramount.
For each model, ProGAN or StyleGAN would be most suitable.
**ProGAN**: Achieved 47% accuracy and 0.471 AUC, near random guessing. Loss plateaued at epoch 15, showing poor learning. Unsuitable due to inability to distinguish classes (Adam, LR=2e-4, batch size=128).
**StyleGAN**: Reached 85% accuracy and 0.930 AUC. Training loss neared zero, but validation loss diverged post-20 epochs, indicating overfitting. Robust detection (85% both classes) but below >99% target (LR=2e-4, batch size=64).
**GJOB-GAN**: Matched 85% accuracy with 0.951 AUC, excelling in fake detection (96% recall) but weaker on real images (73% recall). High validation loss suggests overfitting. Strong but biased (dense layers, dropout=0.25).
**EfficientNetV2**: Efficiency-focused with lower accuracy and slower inference (105.15 ms). Least suited for high-precision needs.

**Scenario 1: Accuracy - Passport Verification**

| Model | Accuracy (%) | AUC | Recommendation | Reasoning |
|---|---|---|---|---|
| EfficientNetV2 | 97% | 0.997 | Best | Highest accuracy; essential for passport validation. |
| StyleGAN | 85% | 0.93 | Acceptable | Good accuracy, but lower than EfficientNetV2. |
| GJOB-GAN | 85% | 0.951 | Acceptable | Good AUC; slight class imbalance. |
| ProGAN | 47% | 0.471 | Unsuitable | Poor accuracy; poses severe security risk. |

## Cost Scenario: Social Media Content Moderation

Scenario 2 involves KlikKlock processing 1 million 30 second videos monthly (12 million annually) for moderation, with cost as the focus. Oracle Cloud compute pricing is $0.021 per OCPU hour for standard VMs. Inference times are StyleGAN/GJOB-GAN at 72 ms, ProGAN at 77 ms, and EfficientNetV2 at 105 ms per video.

| Model | Inference Time (ms) | Total Processing Time (seconds) | Total Processing Time (hours) | Annual Cost (USD) |
|---|---|---|---|---|
| StyleGAN | 72 | 864000 | 240 | 5040 |
| GJOB-GAN | 72 | 864000 | 240 | 5040 |
| ProGAN | 77 | 924000 | 257 | 5397 |
| EfficientNetV2 | 105 | 1260000 | 350 | 7350 |

**EfficientNetV2**: Highest inference time significantly increases operational cost (~45% higher than StyleGAN). Not recommended given KlikKlock's cost sensitivity.
**ProGAN**: Offers moderate computational efficiency with good accuracy balance, making it acceptable though not optimal.
**StyleGAN**: Delivers the lowest inference cost (US$20.05 per million images), ideal for high-volume moderation tasks requiring cost optimisation.
**GJOB-GAN**: Similar cost-effectiveness to StyleGAN, slightly higher (~4.6%), but acceptable where minor accuracy differences are tolerable.

Our recommendation is therefore StyleGAN (lowest cost per million inferences, decent accuracy trade-off. This could be further improved with a larger dataset and more intensive compute allocations to improve hyperparamter selection .

## Latency Scenario: Real-Time Livestream Moderation

For livestream moderation, the key metric is ultra-low latency (<100 ms). Immediate action prevents viral spread of harmful content, safeguarding user trust and regulatory compliance.
**EfficientNetV2**: At 105.15 ms, EfficientNetV2 marginally exceeds the acceptable real-time latency threshold, creating potential exposure risks and non-compliance with regulatory demands (EU content moderation rules).
**ProGAN**: Moderate inference time (77.26 ms) fits well within the real-time responsiveness threshold, effectively balancing image detail processing and latency.
**StyleGAN**: Achieves the fastest inference speed (72.17 ms), best suited to live moderation scenarios requiring instantaneous moderation. Its design inherently supports quick, high-quality discrimination, minimising exposure to harmful real-time content.
**GJOB-GAN**: Similar latency performance (75.53 ms), sufficient for real-time moderation but marginally slower than StyleGAN. Acceptable if slightly relaxed latency is permissible, but inferior in speed-critical contexts.

Therefore, we recommend StyleGAN due to it having lowest latency. This ensures compliance and optimal real-time responsiveness.