AI-Generated Image Detection

DOU, Mingze
mdouab@connect.ust.hk

XU, Zeling
ZelingXu@ieee.org

ZHANG, Shengqi
szhanggd@connect.ust.hk

SHEN, Wenyi
wshenah@connect.ust.hk

1. Introduction

The rapid proliferation of highly realistic AI-generated images poses a serious threat to media integrity and public trust [1]. Existing detectors are often data-hungry, computationally intensive, and exhibit poor generalization to unseen generative models. Compounding the challenge, fast-paced model evolution quickly invalidates historical training data, while collecting diverse samples across generators and prompts remains costly. To address these limitations, we propose a lightweight, data-efficient detector that achieves high accuracy with minimal training data and strong generalization to novel generators.

2. Team Members and Responsibilities

Name	Responsibilities	Proportion
DOU, Mingze	ViT Report writing	25%
XU, Zeling	SPAI Report writing	25%
ZHANG, Shengqi	ResNet50 Report writing	25%
SHEN, Wenyi	NPR Report writing	25%

3. Project Timeline

- **Week 8**: Project Analysis and Proposal — Survey of detection methods
- **Week 9**: Dataset selection and preprocessing
- **Week 10**: Implementation and comparative discussion of backbone models — ViT, ResNet50, and SPAI.
- **Week 11**: Implementation and optimization of the NPR module.
- **Week 12**: Code refactoring, integration and preparation for the final presentation.
- **Week 13**: Technical report writing and reflective project summary.

4. Dataset Overview

We adopt `Tiny-GenImage`, a lightweight subset sampled from the million-scale AI-generated image detection benchmark dataset `GenImage`. `Tiny-GenImage` retains images generated by all seven generative models: ADM, GLIDE, BigGAN, VQDM, Stable Diffusion v1.4 (SDv1.4), Midjourney, and WuKong. For each generative model, the dataset contains exactly 4,000 training images and 1,000 validation images, ensuring balanced representation across different generation mechanisms. Moreover, the number of real images is equal to the total number of AI-generated images, resulting in a perfectly balanced binary classification setup between real and fake categories.

5. Backbone Network Selection

5.1. Model Introductions

5.1.1. Vision Transformer

The Vision Transformer (ViT) partitions an input image into fixed-size patches, embeds them and processes the sequence through Transformer layers.

We consider the following variants, all of which share the same lightweight MLP classification head (see Table 1):

ViT: We adopt the google's vit-base with 16×16 patches and a 768-dimensional embedding. The backbone is frozen during fine-tuning.

ViT+CNN: To compensate for ViT's lack of local inductive bias, a lightweight 2-layer CNN (with BN and ReLU after each layer) is prepended to the baseline model. Its globally averaged output is linearly fused with the [CLS] token via two trainable scalars before the shared MLP head.

Swin Transformer: To address ViT's lack of local inductive bias and multi-scale representation, Microsoft's Swin-Tiny is adopted, featuring 4 hierarchical stages and 12 transformer blocks with a base dimension of 96.

CvT: Microsoft's CvT-13 is employed, a 13-layer model with stage-wise embedding dimensions [64, 192, 384], designed to integrate CNN-style local perception into Transformer layers.

Layer	Configuration
1	Linear(768,512) + BN + GELU + Dropout(0.3)
2	Linear(512,256) + BN + GELU + Dropout(0.3)
3	Linear(256,2) + Softmax

Table 1. Structure of classification head.

5.1.2. ResNet50

ResNet50 consists of 50 layers and employs a bottleneck architecture to construct a four-stage hierarchical feature extractor; its skip connections effectively overcome the degradation problem in deep networks.

ResNet50 + Bayesian Linear Classifier: Traditional CNNs use fixed weights, limiting predictive uncertainty and prone to overfitting; replacing the classifier with a Bayesian linear layer—whose weights follow $\mathcal{N}(\mu, \sigma^2)$ and are trained via ELBO (reconstruction loss + KL divergence)—enables uncertainty-aware inference through Monte Carlo sampling.

ResNet50 + Spatial Attention + Bayesian Linear Classifier: A spatial attention module is inserted after the ResNet50 backbone, which compresses the 2048-channel feature map into two spatial descriptors via average and max pooling, then applies a 7×7 convolution to produce an attention mask for feature reweighting

Variant	Pipeline
ResNet + BNN	ResNet $\rightarrow$ Global AvgPool $\rightarrow$ BNN
ResNet + SA + BNN	ResNet $\rightarrow$ SA $\rightarrow$ Avg Pool $\rightarrow$ BNN

Table 2. Architectures of the ResNet50 variants. SA denotes spatial attention.

5.1.3. SPAI

SPAI leverages the generative-model-independent spectral distribution of real images, using only real data for self-supervised training by randomly masking high- and low-frequency components. It introduces spectral reconstruction similarity—comparing features of original, low-pass, and high-pass images—to cast detection as an out-of-distribution problem. For arbitrary resolutions, a spectral context attention mechanism processes image blocks independently and fuses global context with linear complexity. Key architectural and hyperparameter choices are summarized in Table 3.

Component	Parameters	Value
ViT Backbone	embed_dim / depth / heads	768 / 12 / 12
	patch / image size	16 / 224
	MLP ratio	4
Freq. Restoration	proj_dim / layers	1024 / 2
	mask radius / dropout	16 / 0.5
Classification Head	input dim	$6 \times 12 + 1024 = 1096$
	MLP ratio / classes	3 / 1
PatchViT	min patches	4

Table 3. SPAI Model Parameter Configuration

5.2. Experimental Results

All models are trained on the `Tiny-GenImage` dataset under comparable settings unless otherwise specified.

ViT Variants Results: All ViT-based models are trained for 50 epochs with Automatic Mixed-Precision (AMP). Performance results are summarized in Table 4.

ViT + CNN achieves the highest overall accuracy (87.46%), with improved recall on AI-generated images (89.03%) compared to ViT (86.20%). Swin attains the highest precision on natural images (91.40%) and CvT shows the lowest recall on AI-generated images (72.70%).

Model	Nature		AI		Accuracy
	Precision	Recall	Precision	Recall	
ViT	0.8812	0.8774	0.8577	0.8620	0.8703
ViT + CNN	0.9016	0.8611	0.8461	0.8903	0.8746
Swin	0.9140	0.8443	0.8332	0.9073	0.8734
CvT	0.7980	0.9243	0.8917	0.7270	0.8332

Table 4. Performance of ViT models.

ResNet50 Variants Results: BNN exhibits inverted behavior: it achieves very high recall on real images but lower fake detection performance, resulting in slightly reduced overall accuracy compared to the original ResNet50. By integrating spatial attention, the model achieves the highest overall accuracy (87.6%) among all variants, maintaining strong real image detection while significantly improving fake detection over BNN alone.

Model	Avg Acc	Real Acc	Fake Acc	Uncertainty
ResNet50	86.3%	∼79%	∼92%	✗
ResNet50 +BNN	81.5%	∼85%	∼81%	✓
ResNet50 +Attn+BNN	**87.6%**	**∼89%**	**∼85%**	✓

Table 5. Performance comparison of ResNet50 variants.

SPAI Results: SPAI is reproduced using a pretrained ViT-B/16 backbone, trained for 9 epochs on 7000 images from the dataset.

SPAI achieves 92.0% average precision and 91.0% AUC, demonstrating strong discriminative capability. Although its overall accuracy is 80.5% under the current setup—limited by dataset scale and training configuration—its strength lies in generalization: the original paper reports a 5.5% absolute AUC gain over state-of-the-art methods across 13 generative models, along with robustness to common perturbations such as JPEG compression, Gaussian blur, and resizing.

Metric	Value	Best Epoch
Overall Accuracy	80.5%	9
Average Precision	92.0%	9
AUC	91.0%	7
Minimum Loss	0.6930	2
Training Time	4023.47s	—

Table 6. SPAI Reproduction Results on tiny_genimage Dataset

5.3. Conclusion

We conduct a systematic evaluation of three mainstream architectures—ViT and its variants, SPAI, and the ResNet50 family—on the `tiny_genimage` dataset, with a focus on the trade-off between accuracy and robustness.

ViT + CNN performs best among ViTs, confirming the value of local texture modeling; in contrast, CvT's poor recall on AI-generated images reveals the insufficiency of local cues alone. Similarly, SPAI demonstrates strong discriminative capability through spectral reconstruction similarity; however, its overall accuracy remains limited, and its large model size leads to high inference overhead.

More importantly, both ViT-based models and SPAI exhibit model complexity that significantly exceeds our lightweight deployment requirements, and their accuracy fails to meet the expected target.

Therefore, we shift our focus to the compact and computationally efficient ResNet architecture. Among ResNet50 variants, the version integrating spatial attention and BNN shows notable improvements in class balance and uncertainty awareness. However,BNN's conservative behavior (bias toward "real" predictions) and data inefficiency make it less suitable under limited training data.

Considering inference efficiency, deployment complexity, and performance stability, we ultimately opt not to adopt the BNN classification head.

In summary, we select an architecture based on ResNet50 as the backbone, augmented with targeted feature enhancement and a standard MLP classification head.

6. Final Solution: NPR

6.1. Introduction of NPR

Traditional approaches to deepfake detection, such as CNNs, have primarily focused on designing detection algorithms, with limited investigation into the architectural characteristics of generator models. NPR [2], in contrast, a novel method analyzing upsampling operations in common generator pipelines can extract generalized artifact representations for forgery detection.

6.2. Theoretical Foundation

6.2.1. Upsampling Artifacts Formation

The generation pipeline in modern synthetic image creation follows a structured process: text encoding → UNet encoder → UNet decoder (including upsampling) → VAE decoder [3] → final image generation [4], shown in Figure 1.

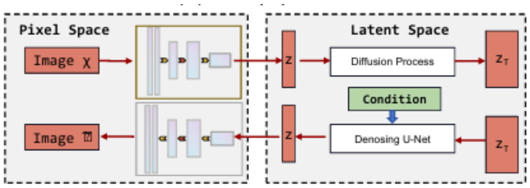

Figure 1. Diffusion generation pipeline highlighting upsampling stages.

Synthetic images are first generated at low resolution and then enlarged via upsampling. For example, under 2× upsampling, a single low-resolution pixel expands into a 2×2 block with initially identical values.

While subsequent convolutional layers process these blocks, their shared origin combined with the translation-invariant nature of convolution causes them to retain unnatural, structured correlations. This creates unique local pixel relationships in the generated image, contrasting sharply with the pixel variations in a real photograph.

6.2.2. Neighboring Pixel Relationships

The core idea of NPR is to reverse-engineer this upsampling process to expose such artifacts. Given an input image x, we:

1. Reduce the original image x to half its size;
2. Immediately enlarge it back to the original dimensions using the same method, resulting in a reconstructed x_{re};
3. Compute the difference map: NPR $= x - x_{\text{re}}$.

For real images, the lossy down-up sampling introduces significant blur, resulting in a large, texture-rich difference map. In contrast, AI-generated images—already containing upsampling artifacts—undergo a reconstruction that partially aligns with their generation pathway, leading to minimal residuals.

This process is visualized in Figure 2.

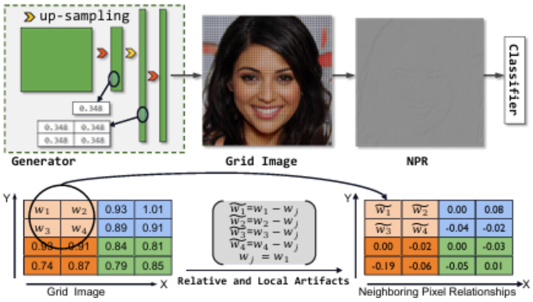
up-sampling
0.348
0.348 0.348
0.348 0.348
Generator
Grid Image
NPR
Classifier
Y
w_1 w_2 0.93 1.01
w_3 w_4 0.89 0.91
0.93 0.91 0.84 0.81
0.74 0.87 0.79 0.85
X
Grid Image
$\widetilde{w_1}=w_1-w_j$
$\widetilde{w_2}=w_2-w_j$
$\widetilde{w_3}=w_3-w_j$
$\widetilde{w_4}=w_4-w_j$
$w_j=w_1$
Relative and Local Artifacts
Y
$\widetilde{w_1}$ $\widetilde{w_2}$ 0.00 0.08
$\widetilde{w_3}$ $\widetilde{w_4}$ -0.04 -0.02
0.00 -0.02 0.00 -0.03
-0.19 -0.06 -0.05 0.01
X
Neighboring Pixel Relationships

Figure 2. Overview of the NPR computation pipeline.

6.3. Limitations of Basic NPR

Despite its theoretical elegance, the basic single-scale NPR formulation faces three practical challenges:

- **Scale mismatch**: Modern generators employ multi-stage upsampling, yet single-scale NPR only captures traces from one specific ratio.
- **Channel imbalance**: Forgery trace intensities vary across different color channels in NPR feature maps, with certain channels containing more discriminative information.

- **Spatial non-uniformity**: Artifactual patterns concentrate in specific regions, but basic NPR treats all spatial locations equally.

6.4. Enhanced NPR Framework

To overcome these limitations, we integrate three complementary components into a unified enhancement framework:

- **Multi-scale NPR fusion**: Multi-scale NPR simultaneously captures upsampling features from both shallow and deep layers. Due to the variation in optimal scale combinations across datasets, a learnable weight W_{scales} is employed to adaptively fuse multi-scale NPR maps through softmax-normalized weighting:

$$\text{NPR}_{\text{fused}} = \sum_{i=1}^{N} w_i \cdot \text{NPR}_{\text{scale}_i}$$

- **Channel attention**: To emphasize informative color channels, we apply channel-wise attention derived from both global average and max pooling:

$$\text{CA}(x) = \sigma\big(f_{\text{MLP}}(\text{GAP}(x)) + f_{\text{MLP}}(\text{GMP}(x))\big),$$

  where σ is the sigmoid function.
- **Spatial attention**: To highlight artifact-rich regions, we generate a spatial mask using concatenated average and max pooling across channels:

$$\text{SA}(x) = \sigma\big(f_{\text{conv}}([\text{AvgPool}(x); \text{MaxPool}(x)])\big).$$

7. Experimental Results and Analysis

7.1. Performance Evaluation of Enhancement Strategies

We conducted ablation experiments on `Tiny-GenImage`, using overall accuracy, average precision (AP), real image accuracy, and generated image accuracy to evaluate the performance of previous improved methods on datasets containing images generated by different models.

Model Version	Overall Acc (%)	AP (%)	Real Acc (%)	Fake Acc (%)
Baseline	95.79	99.22	95.86	95.71
Multi-scale NPR	95.76	99.18	95.40	96.11
Channel Attention	96.51	99.47	95.74	97.29
Spatial Attention	95.34	99.15	98.14	92.54
Final Model	**97.06**	**99.62**	96.03	**98.09**

Table 7. Comparative Performance of Different Enhancement Approaches

The multi-scale NPR approach yields a marginal gain in fake detection accuracy (96.11% vs. 95.71%) but slightly reduces real image accuracy (95.40% vs. 95.86%), likely due to increased feature complexity causing occasional misinterpretation of natural high-frequency textures in real images.

Channel attention emerges as the most effective single enhancement, improving all metrics—particularly fake detection accuracy by 1.58 percentage points (97.29% vs. 95.71%). This stems from its ability to adaptively weight color channels based on generator-specific artifact patterns.

Spatial attention exhibits an imbalanced profile: it achieves strong real image detection (98.14%) but underperforms on fakes (92.54%). This suggests that the method prioritizes visually salient anomalies—specifically, semantically meaningful regions—while overlooking the subtle, distributed upsampling artifacts commonly found in modern generators.

7.2. Best Model Performance Analysis

The integration of multi-scale NPR with channel attention yielded our best-performing model, achieving state-of-the-art results across all evaluation metrics. This combination effectively leverages the complementary strengths of both approaches: multi-scale NPR provides comprehensive artifact coverage across different upsampling patterns, while channel attention ensures efficient utilization of the most discriminative features.

The final model demonstrates remarkable consistency across diverse generative architectures, as shown in the radar chart 3.

The improved model achieves 98.00% accuracy and 99.91% AP, representing a significant improve-

Table 8. Detailed Performance of Final Model Across Different Generators

Generator	Acc (%)	AP (%)	Real (%)	Fake (%)	Samples
adm	97.20	99.78	95.80	98.60	1000
biggan	96.80	99.05	96.40	97.20	1000
glide	97.40	99.87	95.20	99.60	1000
midjourney	95.90	99.26	97.00	94.80	1000
sdv5	97.40	99.83	95.00	99.80	1000
vqdm	98.00	99.91	96.40	99.60	1000
wukong	96.70	99.68	96.40	97.00	1000

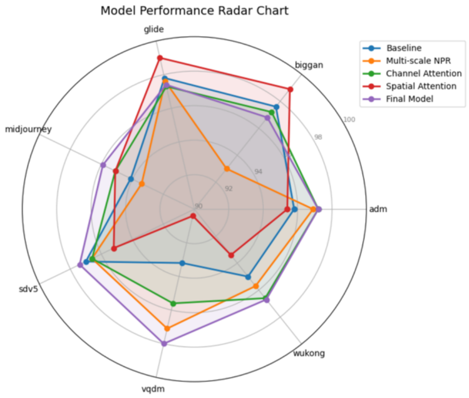

Figure 3. Model Performance Radar Chart

ment over baseline methods. This success can be attributed to the model's ability to capture the unique artifacts characteristic of VQ-based generation pipelines through the multi-scale NPR framework.

The performance on GLIDE (97.40% accuracy, 99.87% AP) and SDv5 (97.40% accuracy, 99.83% AP) underscores the method's effectiveness on diffusion-based generators. These models employ complex upsampling schedules in their denoising pipelines, and our multi-scale approach successfully captures these artifacts.

Midjourney presents the most challenging case (95.90% accuracy), with a noticeable discrepancy between real and fake detection rates (97.00% vs 94.80%). Nevertheless, the model maintains strong performance with 99.26% AP.

The overall performance reveals several key insights: first, the method exhibits slightly higher fake detection accuracy compared to real detection, suggesting that upsampling artifacts provide more reliable information for identifying synthetic content. Second, the high AP scores (all exceeding 99%) indicate excellent ranking capability and detection confidence in all generator types.

The above comprehensive evaluation experiments demonstrates that analyzing upsampling operations through multi-scale NPR with channel attention provides a robust foundation for generalizable deepfake detection. It transcends many generative architectures and training methodologies.

References

[1] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 56(1):1–40, 2023. 1

[2] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 4

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 4

All codes are publicly available at: https://github.com/CVLabWorks/CVIMGDetection.git