Time Step Generating: A Universal Synthesized Deepfake Image Detector

Ziyue Zeng*
Waseda University
Tokyo, Japan
zengziyue@fuji.waseda.jp

Yupei Guo*
Tokyo University of Science
Tokyo, Japan
4624511@ed.tus.ac.jp

Haoyuan Liu
Waseda University
Tokyo, Japan
liuhaoyuan@akane.waseda.jp

Dingjie Peng
Waseda University
Tokyo, Japan
kefipher9013@asagi.waseda.jp

Hiroshi Watanabe
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

Abstract

The rise of high-fidelity text-to-image diffusion models has made synthetic images increasingly indistinguishable from real ones, posing serious threats in digital security and media integrity. Existing detection methods often rely on reconstruction-based pipelines, which are computationally expensive and brittle on out-of-distribution data. We propose Time Step Generating (TSG), a universal synthetic image detector that leverages a pre-trained diffusion model as a feature extractor. By inputting images at a fixed diffusion timestep, TSG captures semantic and structural differences in noise prediction behavior between real and generated images — all within a single forward pass, enabling lightweight and effective classification. To eliminate the reliance on the manually chosen timestep hyperparameter, we further introduce TSG++, an enhanced version that consolidates multi-timestep diffusion features through lightweight fine-tuning. TSG++ learns to align features across all timesteps, producing a unified representation that improves both robustness and generalization without additional inference cost. Experiments on GenImage and challenging multimedia datasets demonstrate that TSG and TSG++ outperform prior methods in both accuracy and efficiency, offering a strong and adaptable solution for diffusion-based synthetic image detection.

CCS Concepts

• **Computing methodologies** → **Computer vision**; • **Security and privacy** → **Human and societal aspects of security and privacy**; • **Information systems** → *Multimedia information systems.*

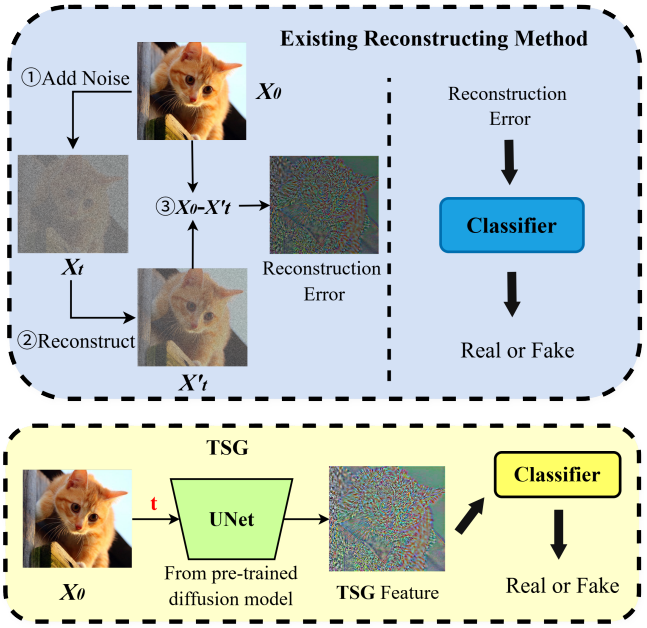

Figure 1: Overview of reconstruction-based methods and our proposed TSG. While reconstruction methods add noise to the original image and then denoise it, TSG directly extracts features at a fixed timestep t using a pre-trained diffusion model, followed by classification.

Keywords

deepfake detection, diffusion models, multimedia forensics, robustness

ACM Reference Format:
Ziyue Zeng, Yupei Guo, Haoyuan Liu, Dingjie Peng, and Hiroshi Watanabe. 2025. Time Step Generating: A Universal Synthesized Deepfake Image Detector. In *ACM Multimedia Asia (MMAsia '25), December 09–12, 2025, Kuala Lumpur, Malaysia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3743093.3770967

1 Introduction

Recently, diffusion models have achieved state-of-the-art performance in the field of image generation. The Denoising Diffusion

*Both authors contributed equally to this research.

MMAsia '25, Kuala Lumpur, Malaysia

ACM ISBN 979-8-4007-2005-5/25/12
https://doi.org/10.1145/3743093.3770967

Probabilistic Models (DDPMs)[15] have introduced a new method for high-quality image generation and have been widely researched. Improvements to diffusion models have focused on multiple aspects, such as accelerating sampling [34, 38, 42], innovate the backbone network[44, 46], improved model framework[7, 31, 33] and optimizing training strategies[16, 25]. Diffusion models have also been investigated for various downstream tasks, including video generation [17], controllable image synthesis [21, 27], and image editing [2, 19]. The proliferation of diffusion model-based technologies in everyday life has raised significant concerns[18] regarding privacy, the dissemination of misleading information, and copyright infringement. Therefore, it is imperative to develop a method for detecting generated images to ensure the integrity of a trustworthy social environment.

The diffusion model represents a significant departure from traditional image generation paradigms, presenting new challenges for forgery detection. In pursuit of a universal detector, recent studies have exploited the reconstruction behavior of diffusion models. **DIRE** [40] hypothesizes that real images are harder for diffusion models to reconstruct, using the pixel-level reconstruction error as a discriminative cue. **LaRE**^{2} [23] improves efficiency by computing the error in the latent space and refining features via error-guided attention. Following this line of work, several methods have further explored reconstruction-based cues while improving generalization and efficiency. **AEROBLADE** [30] proposes a lightweight, training-free approach that leverages only the autoencoder of latent diffusion models, exploiting the mismatch between real images and the model's latent manifold without invoking the diffusion steps. **FIRE** [6] formulates a feature inconsistency score by comparing pre- and post-denoising features, achieving robust detection across both known and unseen generators. However, existing methods either rely on reconstruction errors based on full diffusion inference, where the added noise may degrade the fine-grained details of the input image, or ignore the core denoising network of the full diffusion model.

To address the inefficiencies and fragility of reconstruction-based methods, we propose Time Step Generating (TSG), a simple yet effective detection paradigm that directly uses the noise prediction behavior from a pre-trained diffusion model. Our method relies on two key insights. First, real images are harder to reconstruct than generated ones, making the predicted noise ϵ highly discriminative. Second, in score-based diffusion, real and fake images exhibit different gradient behaviors due to their positions relative to the data manifold. Based on this, TSG extracts features from a pre-trained U-Net at a fixed timestep and classifies them using a lightweight network. This simple and efficient pipeline generalizes well across datasets and generators. An overview is shown in Figure 1.

Through experiments, we observe that features extracted at different timesteps t, especially near the end of the generation process, exhibit varied but useful discriminative patterns between real and synthetic images. Although TSG is efficient and effective, its performance can vary with the choice of timestep t, which limits its generalization across datasets and generative models. To mitigate this, we introduce **TSG++**, an enhanced version that consolidates features from multiple timesteps through a lightweight alignment module. This allows the model to leverage richer multi-scale cues without increasing inference cost.

To evaluate our method, we use the GenImage benchmark [48], which contains 8 subsets generated by different image synthesis models, each with approximately 300,000 samples. On this large-scale dataset, we compare TSG with existing detection baselines, achieving substantial gains in both accuracy and efficiency. Compared to DIRE, our method is 10× faster; compared to LaRE^2, it achieves significantly better accuracy. To further assess generalizability, we conduct experiments on *WildRF* [3], a challenging benchmark with synthetic images collected from real-world multimedia platforms such as *Reddit*, *X* (formerly *Twitter*), and *Facebook*. Due to limited samples and strong distribution shifts, we adopt **TSG++** in this setting and compare it with state-of-the-art deepfake detection methods. Results show that TSG++ maintains competitive or superior performance under these real-world constraints.

The contributions of our work are as follows:

(1) We propose **TSG**, a novel detection framework that leverages noise prediction features from a pre-trained diffusion model, avoiding costly reconstruction.
(2) We introduce **TSG++**, which aligns multi-timestep features into a unified representation, enhancing robustness and eliminating the need for manual timestep selection.
(3) We demonstrate that our method achieves state-of-the-art performance and faster inference than prior approaches across both controlled and real-world datasets

2 RealtedWork

2.1 Image Generation

Early image generation methods such as GANs [8] and VAEs [20] laid the foundation for deep generative models, but often suffer from training instability and limited image quality. Diffusion models, particularly DDPMs [15], have emerged as a more stable and expressive alternative by generating images through iterative denoising from noise. Recent models like Stable Diffusion [31] and PixArt-α [5] achieve state-of-the-art results in high-resolution image synthesis. Meanwhile, autoregressive transformer-based generators [37] are gaining attention for their flexibility and scalability in image and video generation.

2.2 Generated Image Detection

In recent years, research on detecting generated images has primarily focused on GAN-based methods [4, 39], with most approaches relying on CNNs to capture subtle artifacts and inconsistencies. Some works target facial forgery detection [14, 22, 47], while others aim for general-purpose detection across generation models [13, 41]. In addition to spatial cues, several studies investigate frequency-based artifacts [36] or Leveraging large vision-language models [43] to support deepfake detection.

With the emergence of diffusion models, traditional CNN-based methods often fail to generalize due to the absence of GAN-specific artifacts. To address this, several recent methods exploit diffusion model behaviors to detect generated images. DIRE [40] uses reconstruction errors from denoising to distinguish real from fake images. LaRE^2 [23] improves efficiency via one-step latent reconstruction with error-guided attention. FIRE [6] compares features before and after denoising to quantify inconsistencies. Ricker et al. [29] analyze frequency-domain mismatches and diffusion residuals to

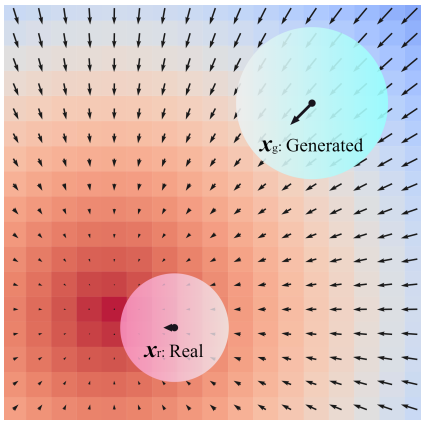

Figure 2: Explain the differences between real and generated samples from the perspective of scores. x_r is the real image's distribution and x_g represents the distribution of generated images. We take the center point of the distribution as an example, the arrow at the center of the distribution represents the estimated score at this point.

detect diffusion-generated images. AEROBLADE [30] bypasses the diffusion process entirely, using only the autoencoder to identify off-manifold samples.

3 Method

This section first reviews the design and rationale of our original Time Step Generating (TSG) method. We then analyze its limitations based on empirical evidence, which motivates the development of an improved framework, TSG++, that unifies timestep-dependent features through lightweight fine-tuning.

3.1 Time Step Generating

We begin with two key observations grounded in the foundations of diffusion models. First, from the perspective of the Denoising Diffusion Probabilistic Model (DDPM), the reverse denoising trajectory progressively refines structural and semantic details from a noisy input. When the timestep t is close to zero—near the end of the denoising process—the U-Net receives inputs with minimal noise and is therefore tasked with recovering fine-grained image features. At this stage, the predicted noise contains rich high-frequency information that is particularly sensitive to subtle discrepancies between real and synthetic images, making it well-suited for downstream classification.

Second, from the Score-Based Diffusion Model viewpoint, the denoising process can be interpreted as gradient-based sampling toward regions of higher data density. Real images lie naturally in such high-density regions, whereas generated images are drawn from learned approximations. This results in inherently different gradient behaviors between real and synthetic samples, especially at low-noise stages. A schematic illustration is shown in Figure 2.

Motivated by these insights, we propose Time Step Generating (TSG), a feature extraction scheme that directly leverages the

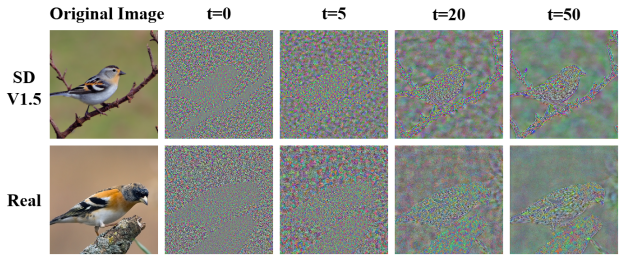

Figure 3: The feature images output by TSG under different conditions of t.

internal noise prediction of a pre-trained diffusion U-Net. Specifically, given an input image, we feed it into the U-Net along with a small timestep t (e.g., 0 or 50), thereby capturing rich semantic features from the model's internal activations. These features are then used as input to a downstream classifier (e.g., ResNet-50 [12]) for real-vs-fake detection. Compared to prior works that add synthetic noise to real images or perform costly inversion-based reconstruction, our one-step approach preserves more authentic detail from real images and avoids the representational degradation caused by artificial corruption.

The formulation of TSG is straightforward and does not rely on reconstruction or inversion. Given an input image I and a small timestep t, we pass them into the pre-trained diffusion model ϵ_θ and extract the predicted noise from the U-Net:

$$F = \epsilon_\theta(I, t), \tag{1}$$

where F denotes the resulting feature representation.

Figure 3 shows TSG feature maps at different timesteps. As t decreases, object boundaries become sharper and the features emphasize structured, high-frequency details. In contrast, at $t = 0$, the features appear noisier but may retain subtle, discriminative cues. This suggests that t effectively controls the granularity of extracted features.

To further validate this, we conduct a quantitative entropy analysis on the *Facebook* subset of the *WildRF* dataset, as shown in Figure 4. Entropy is computed from grayscale histograms using:

$$H = -\sum_{i=0}^{255} p_i \log_2 p_i, \tag{2}$$

where p_i is the probability of grayscale value i. The results show that real and fake images follow distinct entropy trends across timesteps. However, no single t consistently offers the best separation.

These findings highlight a key limitation of TSG: relying on a fixed timestep may fail to capture the full range of discriminative signals. This motivates our enhanced version, TSG++, which consolidates features across all timesteps into a unified representation. We detail its design in the next section.

3.2 Consolidate Timestep-Dependent Features

To overcome the sensitivity of TSG to the choice of timestep, we aim to build a more robust representation that integrates multi-timestep information. Inspired by CleanDIFT [35], we enhance TSG with a fine-tuning strategy that consolidates features across

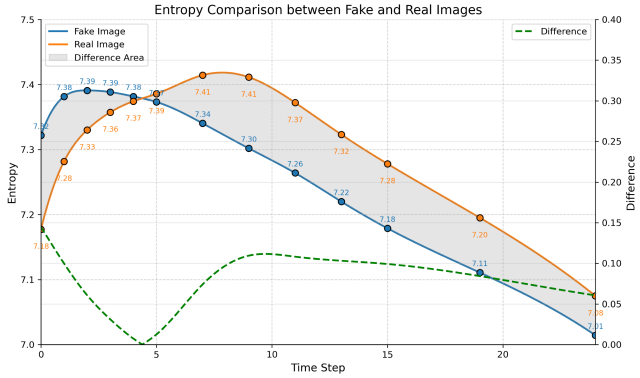

Figure 4: Image entropy of TSG features for real/generated images from the multimedia-based dataset under different time step t conditions.

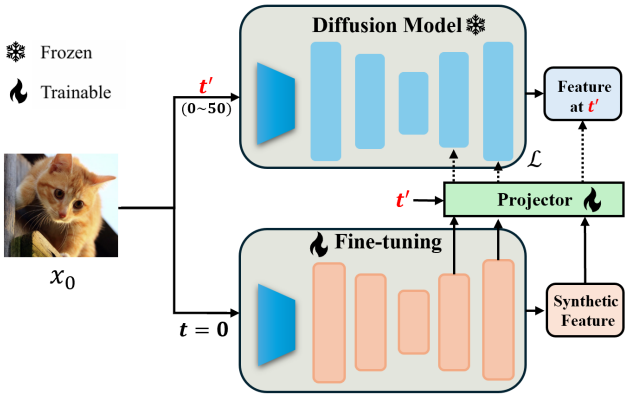

Figure 5: Overview of the TSG++ training framework. A clean image x_0 is simultaneously fed into a frozen diffusion model at a randomly sampled timestep $t' \in [0, 50]$ and a trainable copy at $t = 0$. The synthetic feature from the fine-tuned model is aligned with the reference feature via a timestep-conditioned projector. Alignment is performed at deeper layers including the output.

timesteps into a unified embedding space. Specifically, we introduce a *timestep-conditioned projector* that aligns intermediate representations extracted from the diffusion U-Net at multiple timesteps into a single, timestep-invariant feature space. Unlike CleanDIFT, which performs feature alignment across all decoder blocks, our approach restricts the alignment to only the last two layers of the U-Net—the penultimate downsampling block and the final output block—which carry the most semantically informative and discriminative cues. Each projector consists of two linear layers with GELU activation and layer normalization, conditioned on the timestep embedding t', and is optimized via a cosine similarity loss between projected features and their multi-timestep counterparts. This selective alignment focuses learning on high-level semantics while maintaining low computational cost and preserving the efficiency advantages of single-pass inference.

TSG++ extends the original TSG by consolidating timestep-dependent features through a CleanDIFT-style alignment process. As illustrated in Figure 5, we introduce a fine-tunable replica of the diffusion backbone that receives clean images at a fixed timestep $t = 0$. In parallel, a frozen diffusion model processes the same input at a randomly sampled timestep $t' \in [0, 50]$. To bridge the semantic gap between features extracted at different timesteps, we employ a learnable projector conditioned on t', and align its output with the frozen model's features via a cosine similarity loss:

$$\mathcal{L} = -\cos\left(\text{proj}(F^{\text{tsg++}}(x_0), t'),\ F^{\text{tsg}}(x_0, t')\right), \tag{3}$$

where $F^{\text{tsg++}}(x_0)$ and $F^{\text{tsg}}(x_0, t')$ denote the synthetic feature and reference feature, respectively.

Compared to prior designs, we shift the alignment target deeper into the network, including the final output features. This encourages the model to learn representations that aggregate both high-frequency structure and fine-grained detail from across the diffusion process. Importantly, both inputs are clean images without added noise, ensuring that the learned features are noise-free and task-relevant.

4 Experiments

4.1 Implementation Details

Our experiments are structured progressively to reflect the motivation and effectiveness of our approach. We first analyze the performance of TSG under different timesteps t, exploring how timestep choice affects feature quality and generalization. Based on these findings, we introduce TSG++, which consolidates timestep-dependent features through lightweight fine-tuning. We then validate its performance on more challenging, real-world multimedia data.

Our implementation is based on the DIRE framework. During feature extraction, input images are resized to 256×256, and features are extracted from the U-Net. For classifier training, inputs are cropped to 224×224, and we adopt ResNet-50 [12] as the classifier. Classifiers are trained separately on each generator subset for cross-model evaluation. For TSG++, we follow the CleanDIFT fine-tuning strategy, but replace the backbone with the ADM U-Net used in TSG to eliminate the influence of VAE components.

We evaluate TSG on the GenImage dataset [48], which includes images generated by five different models: *BigGAN* [1], *VQDM* [11], *SD V1.5* [31], *ADM* [7], and *Wukong* [10]. The last four are diffusion-based. Each subset contains approximately 330k images and is split into training and validation sets following the original protocol. To further highlight the generalization ability of different methods, we evaluate TSG++ on the real-world multimedia dataset WildRF [3].

4.2 Comparison with Baselines

We evaluate the generalization ability of our method across different generators in the GenImage dataset. Specifically, we follow the cross-validation setup from previous work, training a classifier on each of the five generator subsets and testing it across all others.

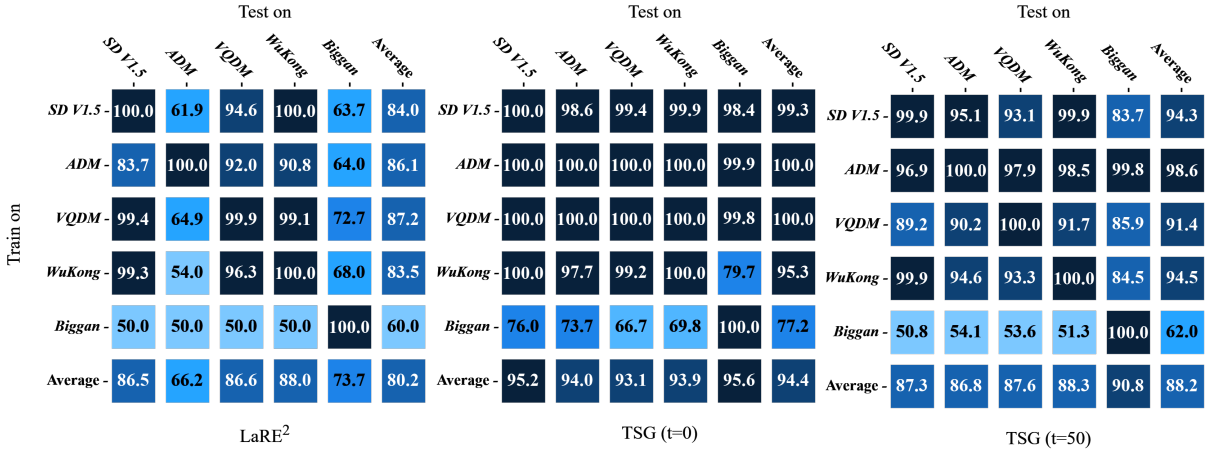

Figure 6: Cross-validation results on various training and testing subsets of GenImage. For each generator, a model is trained and evaluated across all five generator subsets. The matrix plots report accuracy for LaRE^2, TSG at $t = 0$, and TSG at $t = 50$.

As shown in Figure 6, TSG is evaluated under two settings: $t = 0$ and $t = 50$, and LaRE^2 is used as the baseline.

Overall, all methods perform well on the diagonal, indicating near-perfect accuracy when training and testing on the same generator. However, performance drops significantly in the off-diagonal cases, especially for LaRE^2, highlighting challenges in generalizing across generators. In contrast, TSG exhibits better generalization, particularly at $t = 0$, where features contain more detailed and discriminative information. For example, the classifier trained on images from *SD V1.5* or *VQDM* generalizes well to *BigGAN*, despite it being a non-diffusion generator. Similarly, training on *ADM* yields high accuracy across other diffusion models.

We also observe that using $t = 50$ slightly reduces detection performance compared to $t = 0$, supporting our earlier claim that finer details are better preserved at lower timesteps. To further understand this difference, we visualize Grad-CAM activation maps in Figure 7. At $t = 0$, the model attends more to fine-grained, localized regions (e.g., texture and edges), while at $t = 50$, attention shifts to broader, semantic regions such as object shapes and contours.

These complementary patterns suggest that different timesteps capture distinct levels of visual information. Relying on a single t may limit the expressiveness of extracted features. This motivates our proposed TSG++, which consolidates timestep-dependent features into a unified representation to improve robustness and accuracy.

4.3 Evaluation on Real-World Dataset

We evaluate the performance of TSG++ and several baseline methods on the WildRF dataset [3], which consists of real and fake images collected from real-world multimedia platforms, including *Reddit*, *X (Twitter)*, and *Facebook*. Unlike curated benchmarks, this dataset is relatively small and lacks prior information about the generation model types, making it particularly challenging.

All models in Table 1 are trained and evaluated under the same conditions to ensure fair comparison. Compared to previous baselines, our TSG-based methods achieve strong overall results. TSG++

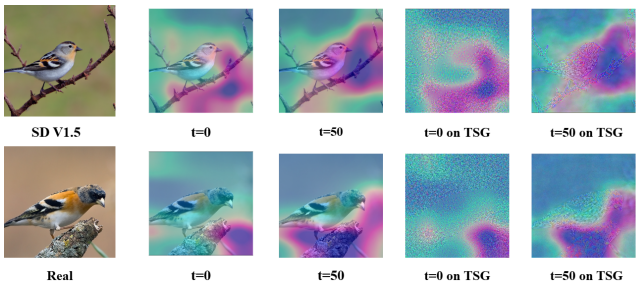

Figure 7: Grad-CAM visualizations of TSG feature maps under different timesteps.

achieves the highest average accuracy (86.2%) and average precision (93.8%) across all subsets. Notably, it shows substantial improvements on the *X* and *Facebook* subsets, demonstrating that integrating timestep-dependent features helps the model generalize better to complex, real-world manipulations.

Interestingly, we observe a slight drop in TSG++ performance on the *Reddit* subset compared to TSG at $t = 50$. We hypothesize that this is due to potential overfitting of the ResNet-50 classifier to domain-specific artifacts or spurious patterns introduced during fine-tuning. Nevertheless, TSG++ maintains competitive performance and outperforms all CNN- and frequency-based methods in both accuracy and AP, further validating its robustness under challenging real-world conditions.

4.4 Efficiency Comparison

To assess inference speed, we compare diffusion-based detection methods on 100 images of size 256×256, with a single RTX A6000 GPU. As shown in Table 2, both **FIRE** and **TSG** achieve significantly faster inference compared to the reconstruction-based baseline **DIRE**.

This efficiency gain mainly comes from bypassing the costly iterative denoising process. While DIRE performs 20-step reverse

Method	Type	*Reddit*		*X (Twitter)*		*Facebook*		Avg	
		ACC	AP	ACC	AP	ACC	AP	ACC	AP
F^3-Net[28]	Frequency-based	85.4	93.1	75.1	92.7	76.9	82.6	79.13	89.47
Effort[45]	VLM-based	73.6	74.8	64.8	84.0	70.9	74.8	69.77	77.87
CLIP[26]	VLM-based	80.8	94.2	78.1	93.1	78.4	90.6	79.10	92.63
CNNDet[39]	CNN-based	75.4	86.8	71.4	84.1	70.6	83.5	72.47	84.80
Xception[32]	CNN-based	86.8	94.4	74.8	93.1	79.1	83.4	80.23	90.30
LaDeDa[3]	CNN-based	91.8	96.0	**83.3**	**92.8**	81.9	92.6	85.67	**93.80**
$Lare^2$[23]	Diffusion	84.4	89.2	74.6	85.5	75.7	85.6	78.23	86.77
FIRE[6]	Frequency+Diffusion	82.4	90.7	70.3	91.3	68.4	77.0	73.70	86.33
TSG (t=0)	Diffusion	91.6	97.5	63.8	64.6	73.8	84.4	76.40	82.17
TSG (t=50)	Diffusion	**92.1**	**97.6**	73.4	80.8	84.7	93.9	83.40	90.77
TSG++	Diffusion	90.2	96.8	82.4	91.7	**86.0**	**94.1**	**86.20**	**93.80**

Table 1: Performance comparison of different methods across datasets from social media platforms. The highest average accuracy and precision are marked in bold.

Method	Parameters			Time (s)
	Batch Size	Sampling Method	Num	
DIRE	20	`ddim20`	100	271.3
FIRE	5	–	100	32.0
TSG	5	–	100	26.3
TSG++	5	–	100	26.5

Table 2: Parameters and time required for diffusion-based methods.

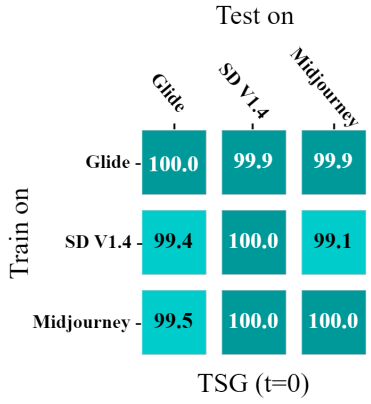

Figure 8: Cross-validation on an unbiased datasets.

sampling per image using the `ddim20` schedule, FIRE and TSG adopt single-pass feature extraction. TSG directly extracts features from the U-Net at a fixed timestep. TSG++ introduces only fine-tuning on the backbone parameters, while retaining the same single-pass inference structure as TSG.

TSG completes inference in just 26.3s, compared to 271.3s for DIRE, demonstrating its suitability for large-scale, real-time deployment.

4.5 Robustness Against Compression

Prior studies on fake image detection have highlighted that image resolution and compression quality can significantly affect classification performance [9]. In particular, lossy compression (e.g., JPEG) may introduce artifacts that inadvertently influence the classifier. To investigate this, we construct three unbiased datasets from the GenImage subsets *Glide* [24], *SD V1.4*, and *Midjourney*, by selecting only images with a JPEG quality factor greater than 96.

We then conduct cross-validation experiments using these high-quality subsets. As shown in Figure 8, the classification accuracies remain consistent with those observed in earlier cross-validation results on unprocessed dataset. This suggests that our TSG feature extraction method is robust to variations in JPEG compression and does not overfit to compression-specific artifacts. Due to computational constraints, we focus on high-quality compression settings in this work and leave broader robustness evaluations, such as low-quality compression, resizing, and cropping, as future work.

5 Conclusion

In this work, we propose TSG, a diffusion-based feature extraction method for detecting generated images. By directly leveraging the U-Net from a pre-trained diffusion model, TSG achieves higher accuracy and faster inference compared to reconstruction-based baselines. It outperforms $LaRE^2$ by 19% in accuracy and runs 10× faster than DIRE. On the WildRF dataset, TSG reaches 83.4% accuracy under challenging real-world conditions. To further enhance performance, we introduce TSG++, which consolidates timestep-dependent features through lightweight fine-tuning. TSG++ achieves state-of-the-art results on multimedia data, demonstrating improved robustness by capturing both fine details and semantic structures.

In future work, we plan to explore the integration of variational autoencoders (VAE) and investigate how latent diffusion models

affect feature extraction performance. As future diffusion architectures evolve, we plan to evaluate the persistence of timestep-dependent discriminative patterns across transformer- and consistency-based generators.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1xsqj09Fm

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

[3] Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. 2024. Real-Time Deepfake Detection in the Real-World. *arXiv preprint arXiv:2406.09398* (2024).

[4] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. 2020. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 103–120.

[5] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=eAKmQPe3m1

[6] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. 2025. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12830–12839.

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 8780–8794. https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[9] Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. 2024. Fake or JPEG? Revealing Common Biases in Generated Image Detection Datasets. *arXiv preprint arXiv:2403.17608* (2024).

[10] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems* 35 (2022), 26418–26431.

[11] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10696–10706.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. 2024. RIGID: A Training-free and Model-Agnostic Framework for Robust AI-Generated Image Detection. *arXiv preprint arXiv:2405.20112* (2024).

[14] Yurika Fujinami Hiroshi Watanabe. 2024. Adversarial Level of Face Images Generated by Prompt-Based Image Coding in Face Recognition System. In *IEEE Global Conference on Consumer Electronics (GCCE2024)*. 332–333.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[16] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.

[18] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. 2022. Countering malicious deepfakes: Survey, battleground, and horizon. *International journal of computer vision* 130, 7 (2022), 1678–1734.

[19] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

[20] Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[21] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. 2025. ControlNet ++: Improving Conditional Controls with Efficient Consistency Feedback. In *European Conference on Computer Vision*. Springer, 129–147.

[22] Jiawei Liang, Siyuan Liang, Aishan Liu, Xiaojun Jia, Junhao Kuang, and Xiaochun Cao. 2024. Poisoned forgery face: Towards backdoor attacks on face forgery detection. *arXiv preprint arXiv:2402.11473* (2024).

[23] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. 2024. LaRE^2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17006–17015.

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[25] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 8162–8171.

[26] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24480–24489.

[27] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. 2024. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070* (2024).

[28] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*. Springer, 86–103.

[29] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. 2022. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571* (2022).

[30] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. 2024. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9130–9140.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[32] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=St1giarCHLP

[35] Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. 2025. Cleandift: Diffusion features without noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 117–127.

[36] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5052–5060.

[37] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. 2025. MAGI-1: Autoregressive Video Generation at Scale. *arXiv preprint arXiv:2505.13211* (2025).

[38] Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K Jha, and Yuchen Liu. 2024. Attention-Driven Training-Free Efficiency Enhancement of Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16080–16089.

[39] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8695–8704.

[40] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. DIRE for Diffusion-Generated Image Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22445–22455.

[41] Alexander Wißmann, Steffen Zeiler, Robert M Nickel, and Dorothea Kolossa. 2024. Whodunit: Detection and Attribution of Synthetic Images by Leveraging Model-specific Fingerprints. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*. 65–72.

[42] Mengfei Xia, Yujun Shen, Changsong Lei, Yu Zhou, Deli Zhao, Ran Yi, Wenping Wang, and Yong-Jin Liu. 2024. Towards More Accurate Diffusion Model Acceleration with A Timestep Tuner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5736–5745.

[43] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2024. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761* (2024).

[44] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. 2024. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8239–8249.

[45] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. 2024. Orthogonal Subspace Decomposition for Generalizable AI-Generated Image Detection. *arXiv preprint arXiv:2411.15633* (2024).

[46] Huijie Zhang, Yifu Lu, Ismail Alkhouri, Saiprasad Ravishankar, Dogyoon Song, and Qing Qu. 2024. Improving Training Efficiency of Diffusion Models via Multi-Stage Framework and Tailored Multi-Decoder Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7372–7381.

[47] Mingxu Zhang, Hongxia Wang, Peisong He, Asad Malik, and Hanqing Liu. 2022. Improving GAN-generated image detection generalization using unsupervised domain adaptation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[48] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2024. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems* 36 (2024).