Spatial-Temporal Forgery Trace based Forgery Image Identification

Yilin Wang[1,2†], Zunlei Feng[1,2,3†], Jiachi Wang[1,2], Hengrui Lou[1,2], Binjia Zhou[1,2],
Jie Lei[4], Mingli Song[1,2], Yijun Bei[1,2*]
[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]School of Software Technology, Zhejiang University
[3]Zhejiang Provincial Key Laboratory of Service Robot
[4]College of Computer Science, Zhejiang University of Technology

Abstract

The rapid development of AIGC technology has enabled highly realistic forged images to deceive human perception, posing serious risks across many areas. Current deepfake image detection methods primarily identify forgeries by extracting handcrafted features, deep features, and frequency-domain features. While these features contain forgery traces, they also include a substantial amount of the image's semantic information, which interferes with the precision and generalization of forgery detection models. To tackle these challenges, this paper introduces a novel forgery image identification method based on the Spatial-Temporal Forgery Trace (STFT). Motivated by the fact that forgery images are more easily fitted to a specific distribution than real images, the STFT method approaches the issue from a forged image distribution modeling perspective, employing generative diffusion models to meticulously capture the temporal distribution of images. It further models the relationship between temporal feature variations and spatially corresponding temporal features, treating them as temporal and spatial forgery traces. Moreover, STFT incorporates frequency-domain features as weighting factors to accelerate the localization of spatio-temporal forgery traces. Experiments demonstrate that by integrating spatial, temporal, and frequency perspectives within the latent space, STFT effectively captures subtle spatio-temporal forgery traces, exhibiting strong robustness and generalizability. It outperforms state-of-the-art methods on major benchmark datasets in the field. The source code for STFT is available at https://github.com/GCLion/STFT.

1. Introduction

With the rapid development of Artificial Intelligence Generated Content (AIGC) technology [5], particularly the maturity of generative models like diffusion models, AIGC has been widely applied in fields such as image and video generation, bringing significant innovation and convenience. However, the widespread application of AIGC has also brought about a series of serious risks. AIGC technology has made it extremely easy to generate realistic images. This has led to the proliferation of false information, malicious tampering, and other harmful activities. Additionally, by generating forged images, malicious attackers may impersonate others to carry out fraud, causing economic losses and a crisis of social trust [51].

Current deepfake image detection methods primarily rely on extracting handcrafted features, deep features, and frequency domain features for forgery identification. Handcrafted feature-based approaches [10, 18] typically extract statistical features such as color histograms, local binary patterns (LBP), and histograms of oriented gradients (HOG) to detect fake images. While these methods may be effective for early deepfake images, they exhibit poor adaptability to emerging deepfake techniques, such as images generated by diffusion models, making them less effective against continuously evolving attack strategies. Methods based on deep features [17, 41] typically utilize deep neural networks such as CNN, ResNet, and EfficientNet to extract high-level features from images for classification and discrimination. While these approaches can improve detection performance to some extent, they often learn the semantic features of the image itself rather than focusing specifically on forgery traces, which affects the generalization ability of the forgery detection model. Frequency domain-based methods [14, 47] extract features from high-frequency or specific frequency bands using frequency domain analysis techniques such as Fourier transform and wavelet transform to detect forgery traces. However, since different AIGC generation methods may use different generation strategies, the stability of frequency domain features is insufficient, which limits the generalization ability of these methods.

The features extracted by the most of aforementioned

*Corresponding author. † Equal contribution to this work.

methods not only contain forgery trace characteristics but also include a large amount of semantic information from the image itself, which interferes with the accuracy and generalization ability of forgery detection models.

Regarding the generative mechanisms of contemporary deep image synthesis models, these models create images by assimilating particular data distributions. However, their gradient-based optimization processes inherently display a propensity for "shortcut learning"—favoring the acquisition of commonly encountered feature distributions while neglecting those that are more infrequent. This phenomenon is driven by the observation that forged image features are easily assimilated, whereas the subtle, rare features inherent in genuine images often evade capture.

This paper adopts a perspective of forgery image distribution modeling, using generative diffusion models to obtain image temporal distribution features. Then, we mine forgery trace features from the temporal distribution feature space that do not contain semantic information interference, aiming to achieve high accuracy and high generalization in image forgery detection. By leveraging the mapped latent temporal distribution of images, this paper elucidates the evolving patterns of temporal distribution features, identifying them as temporal forgery indicators. In the spatial dimension, based on the hypothesis that forgery images tend to model the approximate semantic parts with either exactly the same or completely different distributions, we extract the correlation between the temporal features of corresponding image regions as spatial forgery traces. Furthermore, generated images typically expose forgery traces more easily in high-frequency regions. We obtain high-frequency information as weights to help the model more quickly identify spatio-temporal forgery traces. This approach provides an innovative perspective for research in the detection of forged images using deep learning models. Our main contributions are as follows:

- We introduce an innovative spatial-temporal forgery trace extraction framework designed for the identification of forged images through latent diffusion distribution. Rather than relying on mixed features, the temporal evolution patterns of individual temporal features and the interrelationships among multiple temporal features are modeled as temporal and spatial forgery traces, enabling robust and precise forgery identification.
- Instead of utilizing frequency features directly, high-frequency components are employed as weighting factors to enhance the forgery detection model's ability to swiftly localize spatio-temporal forgery traces.
- Comprehensive experiments demonstrate that the proposed method surpasses the limitations of prior approaches in feature extraction, and exhibits exceptional robustness and generalization capabilities against various forgery techniques.

2. Related Work

2.1. Image Forgery Technology

The field of image generation was initially led by technologies based on Generative Adversarial Networks (GANs) and autoregressive models. Through the continuous evolution of these technologies, milestone products such as BigGAN [3] and DALL·E [13] emerged. However, these methods face certain limitations in content generation control, with stability and complexity issues being particularly prominent. This has led to the emergence of a new type of image generation method—the diffusion model. The Denoising Diffusion Probabilistic Model (DDPM) [19], inspired by nonequilibrium thermodynamics, was born and demonstrated excellent generative performance, further driving in-depth research on diffusion models. Subsequently, researchers have focused on optimizing its structure and improving sampling efficiency. The Latent Diffusion Model (LDM) [38] applies the diffusion process in the latent space, not only improving model efficiency but also enhancing the ability to generate text-conditioned images through a cross-attention mechanism. LDM has become a significant advancement in the field of image generation, leading to the development of well-known generative models such as Stable Diffusion [38] and LCM-RoLA [27].

2.2. Image Forgery Detection

For traditional Photoshop-based forgery methods, early research primarily relied on extracting handcrafted features, deep features, and frequency-domain features for forgery identification. For deepfake images, different approaches have been explored based on distinct feature perspectives. Below, we introduce related works from three aspects: temporal, spatial, and frequency domains.

In forgery detection methods based on spatial features, Sha et al. [39] analyze unique characteristics of generated images and perform model attribution, aiding in source identification and combating the misuse of forged images. The method in [45] detects forged images by comparing the input image with its reconstruction from a diffusion model, demonstrating strong performance in detecting forgeries under unknown diffusion models and various perturbations. Overall, spatial-based forgery detection methods mainly extract image features using deep models. However, these extracted features often mix semantic information with forgery traces, where the presence of semantic features increases the learning difficulty and reduces detection accuracy, ultimately leading to poor generalization performance of the detection models.

In forgery detection methods based on temporal features, [29] effectively identifies images generated by diffusion models by leveraging the deterministic inversion and denoising computation errors of diffusion models, combined

with statistical and neural network techniques. Similarly, [9] analyzes forensic features of GAN-generated images, which can also be used to evaluate the performance of existing GAN detectors on diffusion model images, further improving detection accuracy. Although forgery detection methods based on temporal features utilize diffusion models to extract temporal characteristics, they still rely on supervised signals to extract mixed features from the aggregated temporal features. Like spatial feature-based forgery detection, these methods suffer from interference with model accuracy and poor generalization ability.

In forgery detection based on image frequency domain features, [46] extracts frequency domain features using the Fourier transform and finds that forged images often exhibit different frequency distributions compared to real images. A convolutional neural network is then used to classify high-frequency features. [42] enhances detection performance by learning forgery-specific features in the frequency domain space and analyzing the abnormal patterns of forged images across different frequency bands. These methods determine authenticity based on the quantity and distribution of high-frequency features. However, due to the diversity of image types, high-frequency features also vary, leading to limited cross-domain generalization.

The methods mentioned above are mostly designed for GAN-generated images. Although Corvi et al. [8, 9, 36] also found that spectral artifacts may exist in diffusion-generated images, the effectiveness of spectral artifact-based forgery detection methods in practice remains limited. To address this issue, researchers have begun to analyzing the characteristics of the diffusion generation process itself. For example, DIRE [45] and De-fake [39] use the reconstruction error of the diffusion model to detect forged images, while LaRE2 [28] and AEROBLADE [37] focus on the reconstruction error in the latent space.

However, these methods are highly dependent on pre-trained reconstruction models and are susceptible to the limitations of the training data characteristics. For example, if the reconstruction model is trained on a specific category (such as cats) but is used to detect forged images from different categories (such as dogs or plants), it may result in misclassification. The core issue here lies in using reconstruction errors as the basis for detection, without fully exploring the characteristics of the diffusion generation mechanism itself, thus affecting the understanding of the forgery image generation patterns.

3. Preliminaries

In this section, we define the key notations of the latent diffusion model and outline its basic principles. In LDM [27], given an image $x \in \mathbb{R}^{H\times W\times 3}$ in the RGB space (H and W denote the height and width of the image x), the encoder $\mathcal{E}$ encodes x into a latent representation $z = \mathcal{E}(x)$, while the decoder $\mathcal{D}$ reconstructs the image from the latent representation, yielding $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$, where ideally $z \in \mathbb{R}^{h\times w\times c}$ (h, w, and c denote the height, width, and channel of the latent representation z). This diffusion model consists of a T-step diffusion process q for $1 \le t \le T$ and a denoising process, which is defined as follows:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I), \tag{1}$$

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)), \tag{2}$$

where z_t is the diffusion result at time step t, β_t is the noise factor at time step t, I is the identity matrix, and μ_θ and Σ_θ represent the mean and covariance matrix of the denoising distribution, respectively. After training the model, step-by-step denoising can be performed starting from the standard normal distribution $z_T \sim \mathcal{N}(0, I)$:

$$z_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(z_t - \beta_t \epsilon_\theta(z_t, t)) + \sigma_t \xi, \tag{3}$$

where $\xi \sim \mathcal{N}(0, I)$ is the random noise used to control sampling quality, and σ_t regulates the randomness of the denoising process. Through repeated iterations, the final representation z_0 is obtained, which is then decoded by $\mathcal{D}$ to reconstruct the synthesized image. During the forward diffusion process, high-frequency components are gradually suppressed by noise. By progressively adding noise in the forward process and denoising in the reverse process, high-frequency components become increasingly compressed. Notably, at high noise levels, the loss of high-frequency details becomes inevitable.

4. Method

As depicted in Figure 1, this method employs a diffusion process to map the image into the latent distribution space, subsequently extracting forgery traces from this latent distribution through a comprehensive spatial-temporal analysis. The overall framework consists of the following three main modules: Temporal Prior Correlation Modeling analyzes the latent space representations at different time steps of the diffusion process to capture their temporal variation characteristics. Spatial Correlation Modeling extracts features from different latent dimensions and computes spatial correlations using a self-attention mechanism to model spatial dependencies. Frequency-Enhanced Attention Mechanism utilizes frequency domain information to guide temporal prior correlation computation and spatial correlation analysis, thereby expediting the localization of forgery traces and enhancing generalization capabilities by disregarding the interference of inconsequential features.

4.1. Image Latent Distribution Modeling

To overcome the limitations of existing methods and gain a deeper understanding of the underlying differences between

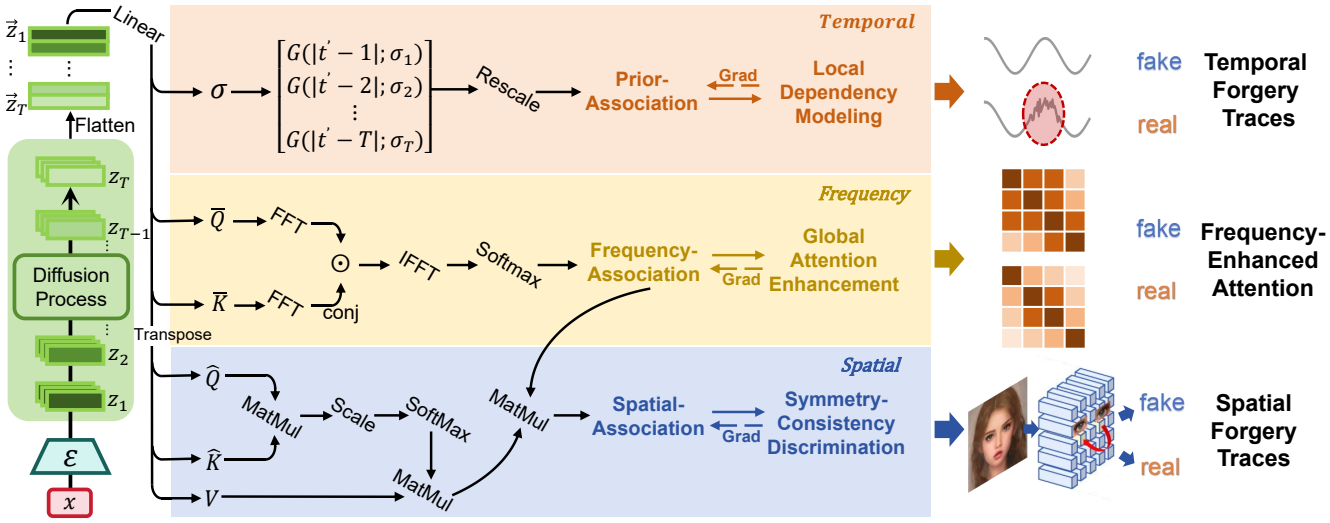


Figure 1. The framework of STFT. It begins by mapping the input image into a latent diffusion distribution space. Subsequently, temporal and spatial forgery traces are extracted by capturing patterns of temporal variation and spatial correlations within the latent diffusion distribution. High-frequency components are utilized as weight to facilitate the rapid localization of forgery traces.

forged and normal images, we propose a forgery detection approach from a distributional perspective. Considering that LDM maps data into a low-dimensional latent space z_t through an encoder $\mathcal{E}$, making the distribution $p(z)$ more compact and easier to model, while the high-dimensional pixel space $p(x)$ remains complex and difficult to model directly, we extract the latent representations $z \in \mathbb{R}^{h\times w\times c}$ during the diffusion process of the latent diffusion model for further analysis.

To facilitate the modeling of distributional changes from a temporal perspective, we reshape the latent representation z_t at each diffusion step into a one-dimensional vector $\vec{z}_t \in \mathbb{R}^{1\times d}$ as sequential data at time step t, where $d = h \times w \times c$, $t \in \{1, 2, \ldots, T\}$. The sequential vectors from step 1 to T are concatenated vertically into a group of tokens $X = \{\vec{z}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$, which serves as input for both temporal and frequency-domain modeling.

4.2. Temporal Prior Association Modeling

Based on the assumption that the latent space distribution of forged images tends to be smoother with weaker temporal variations, we draw inspiration from the unsupervised temporal anomaly representation learning [48] and employ a trainable Gaussian kernel to compute prior associations between time steps. We introduce an adjustable scale parameter σ to adapt to different temporal sequence patterns. The unimodal property of the Gaussian kernel enhances the association between local time steps, making the non-stationary characteristics of real images more prominent, while the smooth evolution pattern of forged images becomes more distinguishable. This improves the temporal modeling capability for forgery trace detection.

For the latent space representations $\{\vec{z}_t\}_{t=1}^{T}$ at different time steps, we treat the diffusion process as a sequential series, concatenate them into $X \in \mathbb{R}^{T\times d}$, and compute prior associations between time steps. Specifically, we define the prior association matrix P as follows:

$$P = \text{Rescale}\left(\frac{1}{\sqrt{2\pi}\sigma_t}\exp\left(-\frac{|t'-t|^2}{2\sigma_t^2}\right)\right), \tag{4}$$

where σ_t represent the learnable scale parameters for the t-th time step. The prior association is computed using a Gaussian kernel, with $G(|t'-t|;\sigma_t) = \frac{1}{\sqrt{2\pi}\sigma_t}\exp(-\frac{|t'-t|^2}{2\sigma_t^2})$ representing the association weight calculated based on the distance $|t'-t|$ between time points t and t'. The function $\text{Rescale}(\cdot)$ normalizes the association weights into a discrete distribution P.

4.3. Spatial Correlation Modeling

Forged images often exhibit abnormal spatial similarities (e.g., excessive symmetry) or inconsistencies (e.g., color or texture mismatches). To address this, we extract features from each dimension of the latent space and apply a self-attention mechanism to compute spatial correlations within the latent space. This helps amplify regions with abnormal similarities or inconsistencies in forged images, thereby enhancing forgery trace detection.

For the feature matrix X, the vertical vector $u_i = X[:, i]$, where $u_i \in \mathbb{R}^{T\times 1}$ and $i \in \{1, 2, 3, \ldots, d\}$, represents the pixel- or region-wise temporal features. Consequently, the spatial token group $X^\top = \{u_i^\top\}_{i=1}^{d} \in \mathbb{R}^{d\times T}$ can be utilized to compute the spatial correlation matrix S as follows:

$$Q, K, V = X^\top W_Q, X^\top W_K, X^\top W_V, \tag{5}$$

$$S = \left(Softmax\left(\frac{QK^\top}{\sqrt{T}} \right) V \right)^\top A, \tag{6}$$

where $\top$ denotes the transpose operation, and Q, K, V along with W_Q, W_K, W_V represent the query, key, and value matrices and their corresponding parameters, respectively. The term A signifies the frequency-enhanced attention weight, which directs the computation of the spatial correlation matrix to emphasize pixel- or region-wise temporal vectors exhibiting significant dynamic evolution patterns. Further elaboration and the corresponding equations for the frequency-enhanced attention matrix A are discussed in the subsequent subsection.

4.4. Frequency-Enhanced Attention Mechanism

In frequency domain analysis, the high-frequency details of forged images tend to be eliminated during gradient optimization, resulting in significant differences in frequency characteristics compared to real images. To address this, we extract the frequency-domain information of latent space features using the Fast Fourier Transform (FFT) and compute frequency-domain correlations to capture cross-scale feature interactions. Since high-frequency information is lost in forged images, the frequency distribution of real and forged images exhibits distinct patterns after conjugate product computation. We then apply the Inverse Fourier Transform (IFFT) to map the correlation back to the spatio-temporal domain, allowing the model to focus on anomalies caused by missing frequency-domain information.

For the group of tokens $X = \{\vec{z}_t\}_{t=1}^T$, the self-attention with query matrix $\overline{Q}$ and key matrix $\overline{K}$ for X is adopted to find the feature temporal similarity of different tokens$\{\vec{z}_t\}_{t=1}^T$. Furthermore, we perform the Fourier Transform on the query matrix $\overline{Q}$ and key matrix $\overline{K}$ to extract frequency-domain features:

$$\widehat{Q} = \mathcal{F}(\overline{Q}), \widehat{K} = \mathcal{F}(\overline{K}), \tag{7}$$

where $\mathcal{F}()$ denotes the Fast Fourier Transform. Next, we compute the conjugate product of the query and key in the frequency domain to obtain cross-scale interaction information C:

$$C = \widehat{Q} \cdot \widehat{K}^*, \tag{8}$$

where "$\cdot$" denotes element-wise multiplication of matrices, $\widehat{K}^*$ represents the complex conjugate, ensuring correct interaction in the frequency domain and highlighting the different frequency distributions of real and forged images. Then, we apply the Inverse Fourier Transform $\mathcal{F}^{-1}()$ to restore the result to the spatio-temporal domain and perform Softmax normalization to ensure that the weights maintain appropriate scales in the spatio-temporal domain, yielding the frequency-enhanced spatio-temporal attention weights:

$$A = \text{Softmax}\left(\frac{\mathcal{F}^{-1}(C)}{\sqrt{T}} \right). \tag{9}$$

Finally, A serves as the frequency-enhanced temporal attention weight to improve generalization ability.

4.5. Learning Strategy

First, we assess the distribution characteristics of the input time series X across multiple levels by computing the association difference ($AssD$) between the prior association P and the frequency-domain association A. Specifically, we employ the Kullback-Leibler divergence (KL divergence) to quantify the discrepancy between two discrete distributions, defined as follows:

$$AssD(P, A) = KL(P[t,:] \| A[t,:]) + KL(A[t,:] \| P[t,:]),$$

where $KL(A[t,:] \| P[t,:])$ denotes the KL divergence, measuring the informational disparity between distributions $A[t,:]$ and $P[t,:]$. Here, $P[t,:]$ and $A[t,:]$ represent the prior association and frequency-domain association distributions at time step t, respectively. The final association difference $AssD(P, A)$ captures the overall variation in association patterns.

For the output of the spatial correlation modeling module, we introduce a global optimization objective based on cross-entropy loss to steer the spatial correlation and frequency-domain enhancement modules in distinguishing between forged and authentic images.

We formulate distinct optimization objectives tailored to different sample types. For forged images, we enforce a deviation between the prior association P and the frequency-domain association A, thereby accentuating the intricate and challenging-to-model characteristics of the real image distribution:

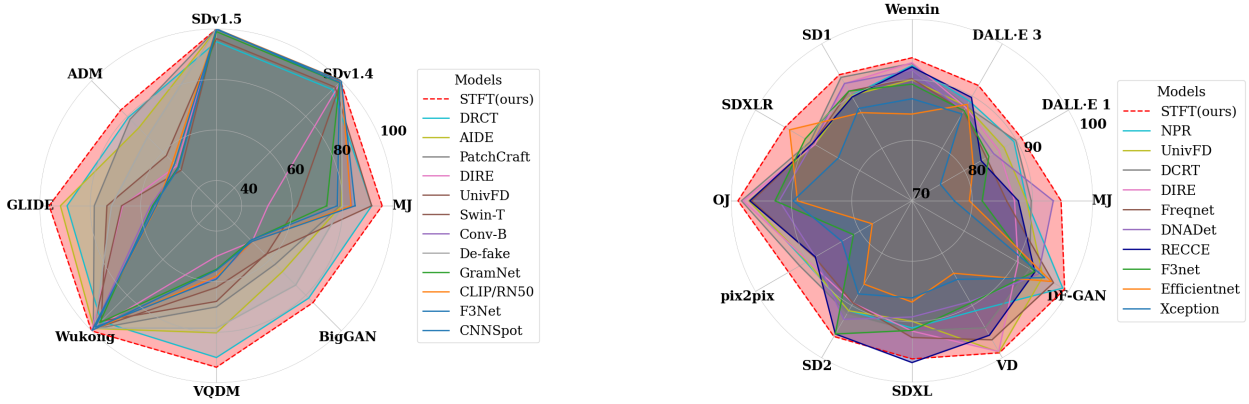
$$\mathcal{L}_{fake} = CE(y, \hat{y}) - \lambda \| AssD(P, A_{fix}) \|, \tag{10}$$

where the first term represents the cross-entropy loss between the classification output $\hat{y}$ of the spatial correlation modeling module and the ground truth labels y, enhancing the module's classification performance. The second term regulates the relationship between the prior association P and the frequency-domain association A using the association difference $AssD()$ for forged images, thereby improving the model's temporal modeling capability. Here, A_{fix} represents the fixed frequency-domain association. The hyperparameter λ modulates the contribution of the $AssD()$ loss.

For real images, we allow the prior association P to closely align with the frequency-domain association A, emphasizing the smoothness of forged images in the frequency domain as follows:

$$\mathcal{L}_{real} = CE(y, \hat{y}) + \lambda \| AssD(P, A_{fix}) \|, \tag{11}$$

where the second term ensures that the prior association P remains distant from the frequency-domain association A



(a) Performance comparisons on Genimage (ACC, %).

(b) Performance comparisons on Deepfacegen (AUC, %).

Figure 2. The radar chart shows the forgery identification performance of STFT in comparison to 12 alternative methods on Genimage (a) and 10 other methods on Deepfacegen (b). Solid lines depict the performance of the comparative methods, while a dashed line signifies the performance of STFT.

for real images, thereby enhancing the model's discriminative capability.

For a batch comprising both forged and real images, in addition to the aforementioned losses $\mathcal{L}_{fake}$ and $\mathcal{L}_{real}$, we ensure that A remains distinct from P, thereby directing greater attention to non-adjacent regions and enhancing the flexibility of global associations:

$$\mathcal{L}_{mix} = CE(y, \hat{y}) - \lambda \|AssD(P_{fix}, A)\|, \qquad (12)$$

where P_{fix} denotes the fixed prior association matrix. This alternating optimization strategy stabilizes the training process and utilizes the frequency-domain association A to guide the model in focusing on more nuanced frequency-domain details, effectively distinguishing between genuine and forged images.

5. Experiments

In the first part of this section, we introduce the detailed experimental setup. Next, in the second part, we evaluate and compare STFT with other forgery detection methods, including SOTA approaches. In the fourth part, we assess the generalization performance of STFT. Finally, we conduct an ablation study on the proposed model.

5.1. Experimental Setup

Evaluation Metrics and Datasets. In alignment with the established comparison standards, we evaluate accuracy (ACC) and area under the curve (AUC) as performance metrics for the GenImage [53] and DeepFaceGen [2] datasets, respectively. The GenImage [53] dataset consists of a total of $2,681,167$ images, including $1,331,167$ real images and $1,350,000$ generated images. These generated images are sourced from eight different generative models: BigGAN [3], GLIDE [30], VQDM [16], Stable Diffusion V1 and V1.5 [38], ADM [12], Midjourney [21], and Wukong [15]. All fake images are generated using the template prompt "photo of CLS," where "CLS" is replaced by one of the $1,000$ labels from ImageNet [11]. The dataset is divided into eight subsets, each containing a portion of real images along with all fake images generated by the aforementioned models. In this study, we follow the official dataset split, using $2,581,167$ images for training and the remaining $100,000$ images for validation.

DeepFaceGen [2] is a facial forgery dataset that includes both video and image modalities. We select face forgery data generated using novel prompt-guided forgery methods. These methods cover 12 diffusion- and autoregressive-based models across text-to-image and image-to-image modalities. The dataset includes face forgeries across various ages, genders, and environments, with generated images sourced from popular models such as Midjourney [21], DALL·E [13], and SDXL [38].

Implementation Details. We downloaded the pre-trained LDM model, fed images into it, and extracted the latent space data. We set the default diffusion steps to 20, obtaining 20 latent representations of size $3 \times 64 \times 64$. Each latent representation was then processed using average pooling to reduce its size to $3 \times 8 \times 8$ and flattened into a one-dimensional vector, ultimately forming sequential data of size 20×192. We set the batch size to 256, the number of channels in the hidden state to 512, and the number of attention heads to 8. The hyperparameter λ was set to 2 for all datasets. We used the ADAM optimizer [23] with an initial learning rate of 1×10^{-4}. All experiments were conducted on an Nvidia A40 GPU.

Method	Year	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN
DIRE [45]	2023	99.5 / 72.6	99.7 / 68.8	99.7 / 68.8	99.2 / 67.9	99.8 / 52.2	99.7 / 68.1	99.9 / 51.9	100 / 50.1
Defake [39]	2023	98.8 / 78.7	99.8 / 81.5	100 / 82.3	98.7 / 82.5	99.8 / 73.1	99.3 / 75.6	99.1 / 82.0	98.7 / 76.6
PatchCraft [52]	2023	90.1 / 81.1	89.5 / 81.3	90.1 / 81.3	90.4 / 79.3	90.6 / 80.4	92.6 / 79.3	90.3 / 74.9	80.9 / 73.7
AIDE [49]	2024	98.9 / 86.0	99.7 / 84.9	99.6 / 84.0	95.3 / 79.1	97.6 / 81.8	98.8 / 80.4	92.3 / 83.0	96.3 / 81.7
DRCT [6]	2024	99.9 / 89.3	95.5 / 86.6	96.8 / 89.1	95.1 / 81.8	96.6 / 83.8	99.6 / 84.4	94.3 / 82.4	98.8 / 84.0
STFT (Ours)	-	**99.4 / 92.3**	**99.6 / 89.1**	**99.3 / 90.4**	**99.5 / 82.5**	**99.8 / 87.0**	**100 / 84.8**	**99.7 / 85.7**	**100 / 85.6**

Table 1. Performance comparison (ACC, %) of various methods. Each value is presented as (Score1 / Score2), where Score1 corresponds to the accuracy achieved when trained and tested on the same subset, and Score2 represents the average accuracy when trained on the current subset and evaluated across other subsets. Bold values highlight the best performance.

5.2. Comparison with State-of-the-arts

GenImage Benchmark. To verify the model's generalization ability across different generative models, this section follows the GenImage evaluation standard and compares our approach with 12 mainstream forgery detection models, including CNNSpot [34], F3Net [33], CLIP/RN50 [34], GramNet [24], De-fake [39], Conv-B [26], SwinT [25], UnivFD [31], DIRE [45], PathCraft [52], AIDE [49], and DRCT [6]. According to the GenImage evaluation protocol, all models were trained on the SDv1.4 subset of GenImage and evaluated on multiple test sets. As shown in Figure 2a, nearly all methods achieve high accuracy on the Wukong, SDv1.4, and SDv1.5 datasets but perform poorly on GLIDE, VQDM, Midjourney, ADM, and BigGAN datasets. Since BigGAN is a non-diffusion model dataset, all methods perform the worst on it. In contrast, our proposed STFT achieves over 89% accuracy on the BigGAN dataset and obtains the highest ACC on 7 out of 8 sub-datasets, demonstrating a significant advantage over existing forgery detection methods (see Appendix Table 4 for complete results).

DeepFaceGen Benchmark. Following the DeepFaceGen evaluation standard, we compare STFT with mainstream generative image detection methods, including Xception [7], EfficientNet-B0 [44], F3-Net [33], RECCE [4], DNADet [50], DIRE [45], DRCT [6], UnivFD [31], NPR [43], and FreqNet [42]. Training and testing were conducted on the same dataset, which includes real images and images generated by 12 different forgery methods such as DALL·E and SD. The dataset was split into training, validation, and test sets with a 7:1:2 ratio. As shown in Figure 2b, STFT outperforms all methods across all 12 subsets, with a particularly notable advantage in distinguishing difficult-to-detect forged images generated by DALL·E 1, DALL·E 3, Midjourney, and Wenxin. Overall, STFT achieves an AUC above 90% for all forgery categories, with an average AUC of 95.21% (detailed results are provided in Appendix Table 5).

Train On	Cat	Horse	Bedroom
Test On	Horse/Bedroom	Cat/Bedroom	Cat/Horse
DIRE [45]	73.01	65.45	55.04
DRCT [6]	74.48	68.78	67.62
STFT (Ours)	**90.32**	**89.70**	**91.51**

Table 2. Generalization performance (AUC,%) of different methods on non-face images.

5.3. Generalization Performance

To evaluate the generalization capability of STFT, we designed two experimental scenarios: first, testing the model when the pre-trained model does not align with the target detection task, and second, assessing its adaptability to unknown forgery methods.

Adaptation to Unknown Forgery Methods. We trained STFT on a single subset of the GenImage dataset and tested it across all eight subsets. As shown in Table 1, we compared STFT with the top four models from the baseline evaluations. STFT outperformed the existing SOTA method in GenImage, DRCT, and achieved the best generalization performance across all eight subsets. When confronted with unseen forgery methods, all models experienced performance degradation in various detection metrics. However, STFT still demonstrated strong generalization ability. The superior generalization of STFT is primarily attributed to its frequency-domain-enhanced spatio-temporal correlation modeling, which captures multi-perspective features of forged images. This enhances the model's adaptability to different types of forgeries, allowing it to effectively recognize anomalous patterns even when facing previously unseen forgery methods.

Generalization Performance on Non-face Images. Reconstruction-error-based forgery detection methods often fail when the test data differs from the reconstruction model's pre-training domain. For instance, applying a model trained on animals to plant images may misclassify real images as fake due to high reconstruction errors. To assess STFT's robustness in such cases, we used pre-trained

Model Variant	Temporal	Spatial	Frequency	AUC(%)
Full Model	✓	✓	✓	92.02
w/o Temporal		✓	✓	86.28
w/o Spatial	✓		✓	79.71
w/o Frequency	✓	✓		89.75

Table 3. Ablation study on different modules.

models from [12] on LSUN Bedroom, Cat, and Horse categories, and compared STFT with DIRE and DRCT. Following the Benchmark 1 protocol on 12 prompt-based forgery datasets, results in Table 2 show that while reconstruction-based methods suffered major performance drops, STFT consistently achieved AUC scores over 90%, highlighting its strong generalization ability.

5.4. Ablation Study

In this section, we conduct ablation experiments to quantify the contribution of each design component in our model and analyze the impact of key hyperparameters. Specifically, we examine the effect of the latent space sampling steps and feature map resolution and the inclusion of the frequency-enhanced attention module.

Effect of Latent Space Sampling Steps and Feature Map Resolution. To investigate the impact of sampling steps and feature resolution in latent space, we evaluate detection performance with sampling steps set to 10, 15, 20, 25, and 30, and test four resolutions by downsampling latent features via average pooling. As shown in Figure 3, the model performs best with 20 sampling steps and a feature resolution of $3 \times 8 \times 8$.

This indicates that 20 steps sufficiently preserve temporal cues while preventing over-smoothing. A $3 \times 8 \times 8$ resolution balances detail retention and generalization. Lower resolutions (e.g., $3 \times 4 \times 4$) lose crucial spatial information, reducing accuracy, while higher resolutions introduce redundancy, affecting robustness. Thus, 20 sampling steps and $3 \times 8 \times 8$ resolution are adopted as default settings for optimal accuracy and efficiency.

Impact Analysis of Model Components. To analyze the contributions of different STFT components, we conduct ablation experiments by sequentially removing the temporal prior correlation module, spatial correlation module, and frequency-enhanced attention module on DALL·E 3 Dataset. The results, presented in Table 3, confirm their complementary roles in forgery detection.

Removing the temporal module significantly weakens the model's ability to capture sequential forgery patterns, leading to a notable performance drop. This aligns with our observation that latent diffusion smooths temporal transitions in forged images. Without spatial correlation modeling, the model struggles to detect local modifications, increasing false positives and reducing robustness. Frequency

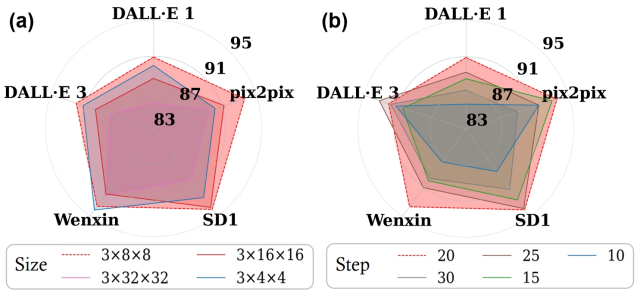


Figure 3. The ablation study on diffusion steps (a) and latent feature map size (b). Solid lines in both cases represent the performance results corresponding to the final adopted values.

features are crucial for detecting high-frequency artifacts in forgeries. Removing this module degrades performance, especially against unknown forgery methods, highlighting its role in generalization.

6. Conclusion

This paper proposes the Spatial-Temporal Forgery Trace (STFT) method for forgery image identification. Leveraging the tendency of forged images to fit specific distributions more easily than real ones, STFT employs generative diffusion models to capture temporal distributions and models relationships between temporal and spatial features as forgery traces. It also integrates frequency-domain features to accelerate trace localization. Experiments show that STFT effectively identifies subtle spatio-temporal traces by combining spatial, temporal, and frequency perspectives, demonstrating strong robustness and generalizability.

Limitations and Future Works. Although the proposed STFT achieves high accuracy and demonstrates commendable generalization capabilities, it is not without its limitations. Firstly, the process of mapping the original image into the latent diffusion distribution is computationally intensive and time-consuming. Additionally, the method appears to exhibit suboptimal performance on GAN-based forgery images, likely due to fundamental differences in the generative mechanisms of GANs and generative diffusion models. In future work, we aim to enhance the efficiency of the proposed method and further extend its generalization ability.

7. Acknowledgments

This work is funded by National Key Research and Development Project (Grant No: 2022YFB2703100), Ant Group through CCF-Ant Research Fund, and Information Technology Center, ZheJiang University.

References

[1] Baidu AI. Wenxin large model for text-to-image generation. *Baidu AI Blog*, 2023. 2

[2] Yijun Bei, Hengrui Lou, Jinsong Geng, Erteng Liu, Lechao Cheng, Jie Song, Mingli Song, and Zunlei Feng. A large-scale universal evaluation benchmark for face forgery detection. *arXiv preprint arXiv:2406.09181*, 2024. 6

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2, 6

[4] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4113–4122, 2022. 7

[5] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023. 1

[6] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024. 7, 2

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 7

[8] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 973–982, 2023. 3

[9] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[10] Edmar RS De Rezende, Guilherme CS Ruppert, Antonio Theophilo, Eric K Tokuda, and Tiago Carvalho. Exposing computer generated images by using deep convolutional neural networks. *Signal Processing: Image Communication*, 66: 113–126, 2018. 1

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6, 8

[13] Tira Nur Fitria. Artificial intelligence (ai) technology in openai chatgpt application: A review of chatgpt in writing english essay. In *ELT Forum: Journal of English Language Teaching*, pages 44–58, 2023. 2, 6

[14] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 1

[15] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022. 6, 2

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 6

[17] Weinan Guan, Wei Wang, Jing Dong, Bo Peng, and Tieniu Tan. Collaborative feature learning for fine-grained facial forgery detection and segmentation. *arXiv preprint arXiv:2304.08078*, 2023. 1

[18] Tahir Hassan, Aras Asaad, Dashti Ali, and Sabah Jassim. Artificial image tampering distorts spatial distribution of texture landmarks and quality characteristics. *arXiv preprint arXiv:2208.02710*, 2022. 1

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[21] Anna Jaruga-Rozdolska. Artificial intelligence as part of future practices in the architect's work: Midjourney generative tool as part of a process of creating an architectural form. *Architectus*, (3 (71):95–104, 2022. 6, 2

[22] Wei et al. Jiang. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16423–16432, 2021. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Zhengzhe 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. 7, 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7, 2

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 7, 2

[27] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 2

[28] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024. 3

[29] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272*, 2023. 2

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 6, 2

[31] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 7, 2

[32] OpenAI. Dall·e 3: Openai's latest text-to-image model. *OpenAI Blog*, 2023. 2

[33] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020. 7, 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7, 2

[35] Aditya et al. Ramesh. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 2

[36] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022. 3

[37] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. 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*, pages 9130–9140, 2024. 3

[38] 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. 2, 6

[39] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pages 3418–3432, 2023. 2, 3, 7

[40] Stability.ai. Stable Diffusion, 2023. 2

[41] Eren Tahir and Mert Bal. Deep image composition meets image forgery. *arXiv preprint arXiv:2404.02897*, 2024. 1

[42] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5052–5060, 2024. 3, 7

[43] 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. 7

[44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7

[45] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 2, 3, 7

[46] Jing Weng. Local frequency analysis for diffusion-generated image detection. In *International Conference on Image Processing and Artificial Intelligence (ICIPAI 2024)*, pages 66–73. SPIE, 2024. 3

[47] Moritz Wolter, Felix Blanke, Raoul Heese, and Jochen Garcke. Wavelet-packets for deepfake image analysis and detection. *Machine Learning*, 111(11):4295–4327, 2022. 1

[48] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*, 2021. 4

[49] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024. 7, 2

[50] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4662–4670, 2022. 7

[51] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023. 1

[52] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. 7, 2

[53] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023. 6