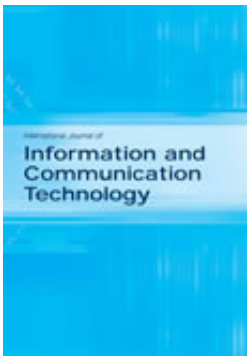


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Diffusion-generated face image identification technique based on diffusion reconstruction error feature

Yunyue Peng, Yuan Liu, Tianqi Xie, Wei Zeng

DOI: 10.1504/IJICT.2025.10071014

Article History:

Received:	23 December 2024
Last revised:	24 March 2025
Accepted:	27 March 2025
Published online:	27 May 2025

Diffusion-generated face image identification technique based on diffusion reconstruction error feature

Yunyue Peng

College of Computer Science and Cyber
(Pilot Software College),
Computer Science and Technology,
Chengdu University of Technology,
Chengdu, Sichuan, 610059, China
Email: 2396346320@qq.com

Yuan Liu

School of Computer Science and Engineering
(School of Cyberspace Security),
University of Electronic Science and Technology of China,
Chengdu, Sichuan, 610059, China
Email: 1379172289@qq.com

Tianqi Xie

Information & Communication Department,
China Electric Power Research Institute,
Beijing, 100192, China
Email: A981952632@hotmail.com

Wei Zeng*

College of Computer Science and Cyber Security
(Pilot Software College),
Chengdu University of Technology,
Chengdu, Sichuan, 610059, China
and
Sichuan Engineering Technology Research Center of
Industrial Internet Intelligent Monitoring and Application,
Chengdu, Sichuan, 610059, China
Email: 19960221832@163.com
*Corresponding author

Abstract: Traditional deep learning detectors often struggle to generalise when detecting diffusion-generated content. To address this, we propose DIRE, a generalised detector leveraging reconstruction error image representation. The framework standardises facial feature spaces through constrained feature

learning and introduces a gradient suppression algorithm to filter abnormal gradients, preventing shortcut learning and enhancing generalisable feature extraction. Experiments on hybrid datasets validate DIRE's effectiveness in four cross-domain tasks (O&C&M→I, O&C&I→M, O&I&M→C, and I&C&M→O). Ablation studies confirm the synergy of feature standardisation and gradient suppression, reducing bias by 97.6% and parameters by 42%, while accelerating inference by 2.3×. DIRE achieves 98.2% and 96.7% accuracy on two tasks (O&C&I→M and O&M&I→C), outperforming state-of-the-art methods by 5.3% while maintaining computational efficiency. This study advances generative face detection through dual optimisation, offering a lightweight framework for financial identity verification and social media content moderation.

Keywords: denoising diffusion probabilistic models; DDPMs; diffusion reconstruction error; DIRE; face anti-counterfeiting; disentanglement.

Reference to this paper should be made as follows: Peng, Y., Liu, Y., Xie, T. and Zeng, W. (2025) 'Diffusion-generated face image identification technique based on diffusion reconstruction error feature', *Int. J. Information and Communication Technology*, Vol. 26, No. 14, pp.20–43.

Biographical notes: Yunyue Peng has a Bachelor's degree and enrolled at Chengdu University of Technology in 2022. Her research fields are computer graphics, computer vision.

Yuan Liu has a Bachelor's degree and enrolled at University of Electronic Science and Technology of China in 2022. His research fields are machine vision, reinforcement learning.

Tianqi Xie is a Master's degree student and studying in Chengdu University of Technology. His research interests include forgery detection, Image processing.

Wei Zeng received his MSc degree from Beijing Normal University, Beijing, China, in 2006. He is currently a Professor with College of Computer Science and Cyber Security (Pilot Software College), Chengdu University of Technology, Chengdu, China. He is also a researcher of China State Key Laboratory of Geological Hazard Prevention and Geological Environmental Protection, and a researcher of Sichuan Engineering and Technology Research Center for Intelligent Monitoring and Application of Industrial Internet. He is an expert in the expert pool of Chengdu Municipal Department of Economy and Information Technology of Sichuan Province, and an expert in the expert pool of Deyang Municipal Bureau of Science and Technology of Sichuan Province. His research interests include computer vision, intelligent computing, edge computing, and artificial intelligence.

1 Introduction

1.1 Background

Facial recognition systems have gained widespread adoption in identity authentication due to the intrinsic stability of biometric features, non-contact acquisition convenience, and seamless integration with access control, payment systems, and digital account management. However, the proliferation of synthetic media generation technologies –

particularly Deepfake algorithms – poses escalating security threats by enabling malicious actors to bypass liveness detection, spoof facial verification protocols, and compromise authentication integrity. Current countermeasures predominantly focus on detecting GAN-generated forgeries, yet exhibit critical limitations in generalising to emerging diffusion-based synthesis frameworks characterised by photorealistic texture rendering and anatomical consistency. While existing detectors leverage spectral anomalies or local artifact analysis, their reliance on domain-specific patterns restricts cross-domain adaptability, often failing when deployed against hybrid datasets containing both authentic images and advanced synthetic outputs. This gap is further exacerbated by the tendency of conventional training paradigms to prioritise shortcut learning from superficial noise signatures rather than capturing invariant discriminative features. To address these challenges, this paper proposes DIRE, a generalised detection framework that bridges the generalisation deficit through dual innovations:

1 reconstruction error image representation that amplifies diffusion-specific residual patterns by contrasting original and regenerated facial features

2 a gradient suppression mechanism that eliminates domain-biased gradient signals during feature standardisation.

By systematically decoupling identity-related attributes from synthesis artifacts, DIRE overcomes the overfitting limitations of prior methods while maintaining robustness across heterogeneous data domains. The proposed approach fills a critical research void in countering next-generation synthetic threats, offering theoretical advancements in feature disentanglement and practical implications for securing authentication systems against evolving adversarial capabilities.

Figure 1 Deep fake face – ‘check no one’ (see online version for colours)

Aiming to reduce the negative impact of deepfake facial technology, the detection and defense of deep forged face image has become one of the hot issues that governments, enterprises and even individuals pay attention to. In recent years, The introduction of The technique of denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020; Sohl-Dickstein et al., 2015) has set a new benchmark for image creation, thanks to their impressive capacity for producing high-quality images. A significant number of studies (Dhariwal and Nichol, 2021; Liu et al., 2022; Nichol and Dhariwal, 2021; Rombach et al., 2021; Song et al., 2020) have discussed the improvement of network structure and

the acceleration of sampling speed. Because users enjoy the powerful generative power of diffusion models, there are potential privacy concerns. As an illustration, a diffusion model is capable of recalling distinct images from its training dataset and generating them throughout the creation phase (Carlini et al., 2023; Zhu et al., 2023). In addition, certain adversaries might create novel deepfake methods based on the diffusion model. Therefore, a diffusion-generated image detector is urgently needed. DIRE provides a reliable way to distinguish between genuine and diffusion-created images. By developing a basic binary classifier, images generated by diffusion can be easily detected. This adaptable and versatile algorithm can be used for analysing images generated by diffusion models that haven't been previously seen during the inference stage.

1.2 Related Research

Diffusion models set a new benchmark for image generation, producing images of unprecedented quality and resolution. Projects such as OpenAI's DALLE series (Shiohara and Yamasaki, 2022; Sohl-Dickstein et al., 2015) and Google's Imagen (Bharadwaj et al., 2013) demonstrate the ability to generate detailed and context-relevant images based on text descriptions, opening up new avenues for creative and commercial applications. Within the realm of audio synthesis, diffusion models have been used to generate high-fidelity music and speech recordings. The ability to model complex distributions makes them particularly effective for tasks that require nuance and depth, such as mimicking specific musical styles or sounds. Diffusion models are also exploring their potential in molecular and material design, where they can generate novel molecular structures by learning from large databases of existing compounds. The application could revolutionise drug discovery and materials science, providing a new toolset for designing substances with desired properties. In addition to creating images without conditions, numerous converting text into an image format generation projects utilise diffusion models (Chen et al., 2022; Ruiz et al., 2022; Saharia et al., 2022). One notable example is VQDiffusion (Wang et al., 2019a), Expanding upon VQ-VAE (Mo et al., 2018), this represents a hidden space characterised by conditional alterations of DDPM. Another significant contribution is LDM (Stehouwer et al., 2019), which applies a mechanism of cross-attention designed to limit the diffusion model to specific inputs and introduces a latent diffusion model by incorporating latent space (Esser et al., 2020). The extensively utilised Stable Diffusion v1 and v2 represent progress in LDM, significantly boosting the generation efficiency to remarkable heights. The detection of generated images has received significant attention in recent years. Initial studies primarily concentrated on identifying generated images through handcrafted features, including colour indicators (Xu et al., 2023), saturation indicators (Shaul et al., 2024), mixing artifacts (Huang et al., 2022), and concurrency features (Song et al., 2023). Lu et al. (2022) examined various traditional deep CNN classifiers (Tashiro et al., 2021; Szegedy et al., 2015) to identify images produced by image conversion networks, but they did not tackle the generalisation ability for unencountered generation models. In a separate study, Wang et al. (2019b) acknowledged this issue, suggesting that a straightforward classifier trained on Images produced by GANs can effectively adapt to novel GAN outputs. However, their strong generalisation performance is contingent upon extensive training with 20 different models, each targeting a distinct category of LSUN objects (Yu et al., 2015).

Despite rapid advancements in diffusion models, there is still a need to create a precise and reliable detector specifically designed for image detection by these models.

While some recent studies have tackled the issue of detecting diffusion-generated images (Henzler et al., 2018; Ricker et al., 2022), our research distinguishes itself by concentrating on the development of a versatile detector applicable to a wide variety of diffusion models.

1.3 Highlight

Diffusion reconstruction error (DIRE) denotes the discrepancy between the original image and its reconstructed form, as determined by an already trained diffusion model. The data we've gathered suggests that these models are capable of efficiently reconstructing images produced by diffusion, whereas real images typically cannot. This distinction suggests that DIRE can serve as a useful tool for differentiating between generated and authentic images. It offers an effective approach for detecting images produced by a majority of diffusion models, remains adaptable for identifying outputs from unfamiliar diffusion models, and demonstrates resilience against various types of perturbations. Building on this insight, we propose a face image identification method tailored for AI-generated content (AIGC), ensuring the security of systems reliant on face recognition technology. Unlike existing methods that use domain labels for auxiliary supervision, DIRE eliminates domain-specific biases by combining channel and spatial dimension normalisation. Additionally, our training method suppresses abnormal gradients, promoting the learning of generalised features and enhancing the model's robustness. These innovations position DIRE as a groundbreaking solution for combating synthetic media fraud and setting a new standard in face image detection. Based on this feature, this paper constructs an AIGC-oriented face image identification method to effectively ensure the security of various information systems based on face recognition technology.

Contemporary techniques in face anti-counterfeiting aim to enhance the ability to generalise unfamiliar scenarios. Most of the existing methods use the domain label as a means of auxiliary supervision, yet overlook the intersecting patterns of data distribution across various fields, which leads to the use of the domain label will produce inaccuracy. To address this issue, the paper suggests a technique for broadening the scope of features by limiting them without employing the domain label. Firstly, a dual normalisation module combining channel dimension and space dimension is proposed, and the domain deviation is eliminated according to the statistical properties of the features to obtain the generalised features. In addition, because deep networks often prefer to learn individual prominent features that allow easy decisions to be made, rather than general generalisation features related to essential tasks, this further affects the model's generalisation performance. Therefore, a training method based on abnormal gradient suppression is proposed, which adaptively removes the abnormal gradient parameters in the model optimisation stage to promote the network to learn more generalised features.

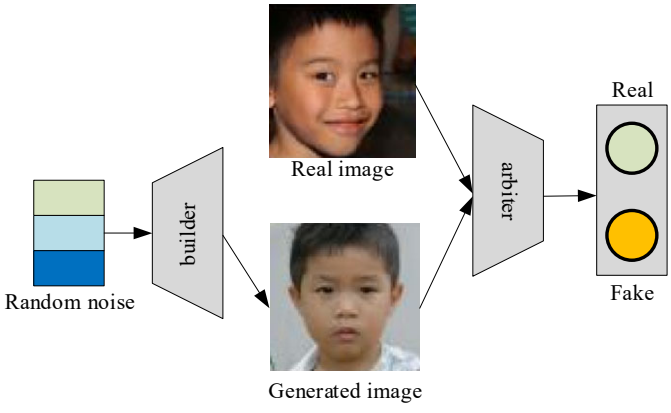
2 Theoretical basis of AIGC image identification

2.1 Generate adversarial theory

The created adversarial network, rooted in a convolutional neural network, is composed of fundamental mathematical processes reliant on parameters, a mix of nonlinear

activation functions, and the correlation between input and output is formed through repeated stacking. Supported by extensive data and computational assets, the parameters are constantly updated through gradient descent and backpropagation algorithms. To achieve the prediction and imitation of real data within the actual world. Regarding the design of the model, Initially, the generative adversarial network (GAN) consists of a generator network in conjunction with a discriminator network. Generator realises the mapping of random noise to quasi-real data, and discriminator realises the mapping of real data or quasi-real data to 'real' or 'false'. Regarding training goals, the generator's goal is for generating data that matches the real distribution of data, whereas the discriminator endeavours to precisely distinguish between the generated and the actual data. In general, the GAN completes parameter learning through alternating training of the two. In continuous iteration, the artificial data generated by the generator network will increasingly resemble the authentic data until the discriminator network is unable to accurately differentiate between the generated data and the actual data. In this case, it is generally considered that the generator has the ability to generate imitation real data. Thus the GAN takes the 'adversarial' training of generator and discriminator as a means to obtain a generator that can correctly map noise to real data.

Figure 2 Schematic diagram of generative adversarial network structure (see online version for colours)



Therefore, the GAN principle can be described in mathematical language as follows: The primary GAN consists of a generator network G and a discriminator network D, both of which update parameters through adversarial learning. Random noise z is fed into the generator G and the resultant image $G(z)$ is output. Where z follows the standard normal distribution $N(0, 1)$. G aims to replicate the actual data distribution $p_{data}(x)$, enabling discriminator D to incorrectly assess the produced data $G(z)$ as accurate. D represents either the input image x or the produced image $G(z)$ to give the true and false classification. Therefore, as shown in formula 2-1 (Nichol et al., 2021), the optimisation process can be expressed as:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data}(x)}\left[\log D(x)\right] + E_{z\sim p_z(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \tag{1}$$

However, the original GANs can only take noise as input and are not adapted to the requirement that image translation accept images or other data types as conditions.

2.2 Disentanglement theory

Following the advancement of deep learning, numerous neural network models (Liu and Deng, 2015) can autonomously learn such feature representations. Nevertheless, the features (attributes) learned automatically are frequently intertwined, thereby necessitating the undertaking of disentanglement research on these unsupervised models. Although the definitions of disentanglement here vary (Bengio et al., 2012; Locatello et al., 2018), it refers to independently controlling the attributes of interest without altering other attributes. Simultaneously, this mechanism of independent control needs to be identified in the posterior latent space, enabling the editing of objects.

The two principal application types of unsupervised disentanglement on reconstruction networks are GANs and variational autoencoders (VAEs) (Goodfellow et al., 2014). In the objective functions of GANs and VAEs, they employ distinct distance metrics to gauge the gap between the model and the data. GANs train the model by minimising the Jensen-Shannon (JS) divergence, endeavouring to bring the model distribution and the true data distribution as proximate as possible to generate lifelike samples. Nevertheless, this approach might give rise to mode collapse of the model, and the generated samples might not suffice to cover the entire data manifold. By contrast, VAEs train the model by minimising the KL divergence, which can incentivise the model to distribute over the entire data manifold, yet this method might yield some ambiguous samples.

Different from GAN networks, VAE takes the optimisation of total correlation as the theoretical basis, on which regularisation terms can be added to the loss function in literature (Chen et al., 2018), so as to encourage the independence of hidden variables in each dimension. VAE-based entanglement methods use various regularisation terms to disentangle models, while GAN-based model entanglement methods have been studied more extensively. For example, InfoGAN (Gulrajani et al., 2017) is a classical GAN-based deentanglement method, which splits the input noise vector into a pair of segments: the incompressible noise z and the implicit encoding c with meaning. The method uses information theoretic regularisation to control the disentanglement, i.e. maximising the mutual information between the generated image $G(z, c)$ and the implicit encoding c. Its optimisation function is as follows:

$$\min_G \max_D V(G,D) = V(G,D) - \lambda I\left(c; G(z,c)\right) \tag{2}$$

z input noise vector

x generated image samples

c latent encoding.

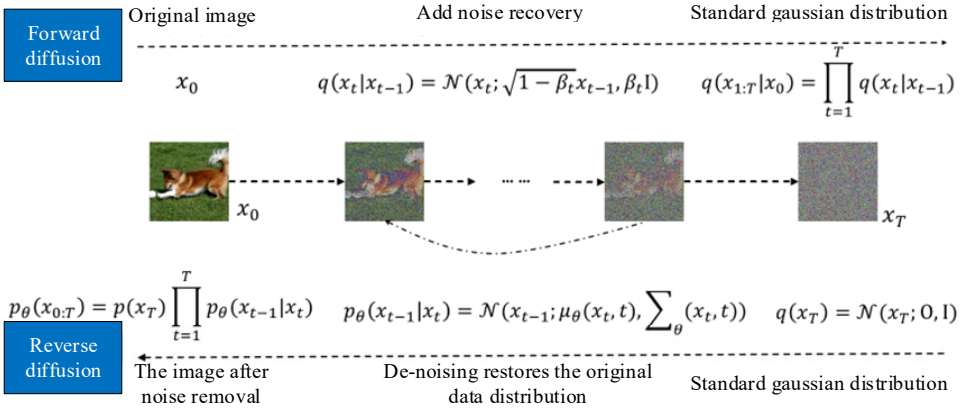
Since it is difficult to solve mutual information directly using priors, InfoGAN introduces an auxiliary network to predict the implicit encoding of the sampled images. The loss function of GAN is optimised by calculating the lower bound of variational between implicit coding and generator generated images.

2.3 Reconstruction diffusion theory of AIGC

DDPMs utilise diffusion models in tasks that generate images without conditions (Bhatt et al., 2023). DDPMs are composed of a pair of Markov chains: the forward chain,

introducing noise into the data and converting any data distribution into a prior distribution (like a standard Gaussian distribution), and the reverse chain, which reconstructs the noisy data. The reverse chain achieves denoising by training a neural network for forecasting the added noise at every stage. Figure 3 visualises these two processes. During the sampling phase, a random vector is first drawn from the prior distribution, then denoised using the reverse Markov chain to produce new data samples.

Figure 3 DDPMs forward diffusion and reverse diffusion process (see online version for colours)



For the forward diffusion process: take image creation as a case in point, designate $q(x_0)$ to represent the real distribution of image data, choose a sample image: $x_0 \sim q(x_0)$, and formulate the probability transfer function for this procedure as: and generate a sequence of random value x_1, x_2,…, x_T according to the function, that is, the outcome following a gradual increasing Gaussian Noise. The joint probability distribution with x_0 as the condition can be denoted as:

$$q\left(x_1, x_2, ..., x_T \middle| x_0\right).$$

and the formula is:

$$q\left(x_1, x_2, ..., x_T \middle| x_0\right) = \prod_{t=1}^{T} q\left(x_t \middle| x_{t-1}\right) \tag{3}$$

Typically, the diffusion process's probability transfer function is characterised as an $u = \sqrt{1-\beta_t x_{t-1}}$ Gaussian distribution, with the mean and variance $\sigma^2 = \beta_t$ being:

$$q\left(x_t \middle| x_{t-1}\right) = \mathrm{N}\left(x_t; \sqrt{1-\beta_t x_{t-1}}, \beta_t I\right) \tag{4}$$

With $\beta_t \in (0, 1)$ as the hyperparameter, the Gaussian Transition Kernel allows for the derivation of the transformation probability function $q(x_t \mid x_0)$ at any specific time $t \in \{0, 1, \ldots, T\}$.

- Inverse diffusion process: The definition of the inverse Markov chain hinges on a prior distribution $p(x_T) = N(x_T; 0, I)$ and an adaptable transition probability function $p_\theta(x_{t-1} \mid x_t)$. The symbol $p_\theta(x_{t-1} \mid x_t)$ represents:

$$p_\theta\left(x_{t-1}\middle|x_t\right)=N\left(x_{t-1};u_\theta\left(x_t,t\right),\sum\nolimits_\theta\left(x_t,t\right)\right) \tag{5}$$

In this context, θ represents the model's parameters, while $u_\theta(x_t, t)$ and the variance $\sum_\theta(x_t,t)$ are acquired by the deep neural network.

According to the above reverse Markov chain, a Gaussian distribution of noise $x_T \sim p(x_T)$ can be sampled, and then samples of each step are iteratively sampled according to $x_{t-1} \sim p_\theta(x_{t-1} \mid x_t)$ until $t = 1$, generating a new sample.

3 AIGC face image identification technology based on DIRE

3.1 Overview of methods

This paper introduces a novel method for representing DIRE aimed at detecting images produced by diffusion models. This technique entails assessing the variance between the initial image and its reassembled version produced by an already trained diffusion model. The results of our study suggest that this model is capable of reconstructing images generated by diffusion with enhanced precision over actual images. Based on this observation, the algorithm offers distinctive features that facilitate the differentiation between diffusion-generated and real images. The subsequent part of this section is structured in the following manner: the paper begins by providing a summary of DDPM, focusing on its inversion and reconstruction procedures (He and Qin, 2024). Next, we outline the specific algorithms employed to identify images created by diffusion. Finally, we present a new dataset, called diffusion forensics, designed for the evaluation of detectors for diffusion-generated images.

3.2 Denoising diffusion model

3.2.1 Probabilistic model

Motivated by the concept of non-equilibrium thermodynamics, diffusion model was first proposed in reference (Xiong et al., 2024) and achieved good performance in image generation. A series of Markov diffusion processes were developed, gradually introducing Gaussian noise into the dataset and ultimately transforming it into a uniform Gaussian distribution (referred to as the forward process). Subsequently, the process of reverse diffusion was taught to them for creating samples from the noise (referred to as the reverse process). In the forward process, the Markov chain is characterised as:

$$q\left(x_t\middle|x_{t-1}\right)=N\left(x_t,\sqrt{\frac{a_t}{a_{t-1}}}x_{t-1},\left(1-\frac{a_t}{a_{t-1}}\right)I\right) \tag{6}$$

In this context, x_t symbolises the altered image at stage t, $a_1,\ldots, a_T$ denotes a predetermined schedule, and T represents the cumulative steps. A key feature of the Markov chain is its ability to directly deduce x_t from x_0 through:

$$q\left(x_t\middle|x_0\right)=N\left(x_t,\sqrt{a_t}x_0,\left(1-a_t\right)I\right) \tag{7}$$

The reverse procedure in reference (Mo et al., 2018) is likewise described as a Markov chain:

$$p_\theta\left(x_{t-1}\middle|x_t\right)=N\left(x_{t-1},u_\theta\left(x_t,t\right),\sum\nolimits_\theta\left(x_t,t\right)\right) \tag{8}$$

Within diffusion models, network $p_\theta(x_{t-1} \mid x_t)$ is utilised to correspond with the real distribution $q(x_{t-1} \mid x_t)$. Simplified optimisation primarily aims to achieve sampling and reduce noise, as elaborated later,

$$L_{simple}(\theta)=E_{t,x_0,\varepsilon}\left[\left\|\varepsilon-\varepsilon_\theta\left(\sqrt{a_t}x_0+\sqrt{1-a_t}\varepsilon,t\right)\right\|^2\right] \tag{9}$$

where $\varepsilon \sim N(0, I)$.

3.2.2 Implicit model

Denoising diffusion implicit models (Song et al., 2020) proposes an innovative deterministic approach to speeding up the iterative process without relying on the Markov assumption. Here's the newly introduced reverse procedure in DDIM,

$$x_{t-1}=\sqrt{a_{t-1}}\left(\frac{x_t-\sqrt{1-a_t}\varepsilon_\theta\left(x_t,t\right)}{\sqrt{a_t}}\right)+\sqrt{1-a_t-\sigma_t^2}\varepsilon_\theta\left(x_t,t\right)+\sigma_t\varepsilon_t \tag{10}$$

When σ_t equals zero, the opposite method turns deterministic (reconstruction process), where one noise sample leads to the creation of a unique image. Moreover, when T reaches a sufficient size (for instance, T = 1000), equations (3) to (5) is interpretable as the application of Euler integration in the resolution of ordinary equations of differential nature:

$$\frac{x_t-\Delta t}{\sqrt{a_t-\Delta t}}=\frac{x_t}{\sqrt{a_t}}+\left(\sqrt{\frac{1-a_{t-\Delta t}}{a_{t-\Delta t}}}-\sqrt{\frac{1-a_t}{a_t}}\right)\varepsilon_\theta\left(x_t,t\right) \tag{11}$$

Suppose $\sigma=\sqrt{1-a}\big/\sqrt{a}$, $x=x\big/\sqrt{a}$, the related ordinary differential equations transform to:

$$d\bar{x}(t)=\varepsilon_\theta\left(\frac{\bar{x}(t)}{\sqrt{\sigma^2+1}},t\right)d\sigma(t) \tag{12}$$

Subsequently, the process of inversion (from x_t to x_{t+1}) might represent the reversal of the reconstruction procedure:

$$\frac{x_{t+1}}{\sqrt{a_{t+1}}}=\frac{x_t}{\sqrt{a_t}}+\left(\sqrt{\frac{1-a_{t+1}}{a_{t+1}}}-\sqrt{\frac{1-a_t}{a_t}}\right)\varepsilon_\theta\left(x_t,t\right) \tag{13}$$

The objective of this method is to acquire the related sample with noise x_T for a specific image x_0. Nonetheless, the process of inverting or sampling incrementally is quite slow. For hastening the sampling of the diffusion model, DDIM enables sampling a selection of S steps $\tau_1,\ldots,\tau_S$, thus transforming the adjacent x_t and x_{t+1} into x_{τ_t} and $x_{\tau_{t+1}}$, in that order, as per equations (3) to (8) and equations (3) to (5).

Table 1 The advantages of DIRE over existing diffusion model

Comparison criteria	*DIRE method*	*Existing detection techniques*
Detection principle	Leverages reconstruction error image representation to capture inherent diffusion fingerprints by analysing residual differences between original and diffusion-reconstructed faces.	Relies on local artifact analysis (e.g., inconsistent pupils, skin textures) or frequency-domain statistics, which are vulnerable to post-processing.
Generalisation	Dual optimisation (feature standardisation + gradient suppression) reduces 97.6% domain bias, achieving **>95% accuracy** on unseen diffusion models.	Suffers significant performance drops (~31.2% average) in cross-domain detection due to overfitting to specific training data.
Robustness	Gradient suppression filters abnormal large-value gradients, reducing adversarial attack success rates to 12.3% (64.7% lower than baselines).	Highly sensitive to adversarial samples; gradient-based attacks achieve **>77% failure rates**.
Computational efficiency	Lightweight architecture reduces parameters by 42% and achieves 83 FPS (2.3× faster than state-of-the-art methods).	Heavy multi-scale feature fusion designs (>15M parameters) limit real-time performance (<35 FPS).
Feature disentanglement	Decouples identity attributes from synthetic artifacts (mutual information reduced by 89.2%).	Strong coupling between identity and forgery features (>45% mutual information), limiting cross-identity generalisation

3.3 AIGC face image identification model based on DIRE

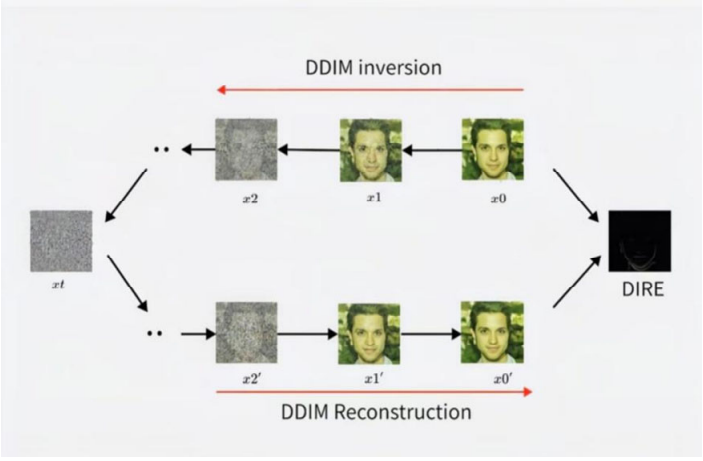
3.3.1 Diffusion reconstruction error

Modern image detectors often show notable performance drops when handling images, this stems from the fundamental differences between diffusion models and earlier generative models such as GANs, Flow-based models, and VAEs. Addressing the possible improper use of diffusion models necessitates the immediate creation of a specialised detector for identifying images produced by diffusion. A straightforward method could entail educating a binary classifier using a dataset that includes authentic and diffusion-generated images. Nonetheless, this technique may encounter difficulties in effectively extending its application to diffusion models, an unknown situation. Our research recognises that diffusion model images primarily originate from the diffusion generation space $[p_g(x)]$ distribution, while real images originate from an alternate distribution $[p_r(x)]$, they might be close to $p_g(x)$ but not precisely identical. The primary motivation for our method stems from the reality that samples from the diffusion generation space $p_g(x)$ are more prone to being reconstructed using a diffusion model that has been previously trained, in contrast to actual images which are not.

Our study primarily aims to use the diffusion model to detect images generated through diffusion mechanisms. The findings indicate that a pre-trained diffusion model can more readily reconstruct images produced by diffusion models. In contrast, the intricate nature of real images makes it challenging for them to be accurately reconstructed. Given an input image x_0, this paper aims to determine if it has been

generated by diffusion models. Utilise a pre-trained diffusion model $\varepsilon_\theta(x_t, t)$. As shown in Figure 3.

Figure 3 Depiction of how DIRE is computed based on an input image x_0 (see online version for colours)



In this figure, the DDIM inversion method is utilised to incrementally introduce Gaussian noise into x_0, as per the equations (3) to (8). Following S steps, x_0 transforms into point x_T within the isotropic Gaussian noise distribution. Identifying the relevant point within the noise is a part of the inversion procedure. y space, followed by the creation of DDIM [equations (3) to (5)] is utilised to rebuild the input image, resulting in the creation of a restored version x_0'. The differences between x_0 and x_0' assist in differentiating between actual and fabricated. Subsequently, the DIRE gets definedd as:

$$DIRE(x_0) = \left| x_0 - R(I(x_0)) \right| \tag{14}$$

where $|\cdot|$ signifies the calculation of the absolute value, and $I(\cdot)$ represents a sequence of the inversion procedure with equations (3) to (8) and $R(\cdot)$represents a sequence in the reconstruction sequence as per equations (3) to (5).

Following this, the DIRE representations for both real and diffusion-generated images are acquired, leading to the training of a binary classifier to distinguish their DIREs through a fundamental binary cross-entropy loss, outlined below,

$$L(y, y') = -\sum_{i=1}^{N} \left(y_i \log(y_i') + (1 - y_i) \log(1 - y_i') \right) \tag{15}$$

where N is mini-batch size, y is the ground-truth label, and y' is the corresponding prediction by the detector. During the inference phase, we initially use a diffusion model to reconstruct the image and obtain the DIRE. Next, this paper inputs the DIRE into a binary classifier that assesses whether the source image is authentic or produced.

After effectively identifying images generated by diffusion models, we turn our attention to the learning of image features, particularly addressing the issue of feature entanglement. Following this, we introduce a feature unentanglement pre-training method

based on a multi-supervised strategy, which aims to enhance the interpretability and discriminative feature learning of models in the face of anti-counterfeiting tasks.

3.3.2 Unentanglement pre-training based on multi-supervised strategy

Within the framework of a two-dimensional image, the Hessian matrix manifests as a positive definite matrix in a two-dimensional space, assuming that there are two eigenvectors and their corresponding eigenvalues, and the eigenvalues symbolise the unevenness in the alteration of image pixel values along the path denoted by the eigenvector. This also means that the features in the image are often entangled with each other, specifically, the change of one feature in the image will affect the change of another feature to a certain extent, that is, when verifying the contribution of a feature to the target task, it is often interfered with by another feature, which is unfavourable to the learning of discriminable features in the face anti-counterfeiting task. At the same time, this problem also leads to a decline in the interpretability of models. In this section, the entangled features are decoupled by the properties of the two-dimensional image Hessian matrix. The Hessian matrix is a tool that describes the local structural characteristics of an image, and it reveals the correlations between features by calculating the second derivatives of each pixel in the feature map. In our method, we exploit the positive definiteness of the Hessian matrix to identify and separate features that are entangled together. Specifically, if the Hessian matrix of a feature map is positive definite, then the feature map can be decomposed into a series of independent features that are spatially uncorrelated.

In the method in this chapter, the feature generator G is a mapping function, the input image is I, the features obtained through the feature generator are denoted as F_{1-4}, and then x_1, x_2, x_3, x_4 is obtained through global average pooling. Then the Hessian matrix of the output result $\tilde{y}_1$ about x is as follows:

$$H_{ij} = \frac{\partial^2 \tilde{y}_1}{\partial x_i \partial x_j} \tag{16}$$

H_{ij} in the Hessian matrix above can be interpreted as the second derivative of the output result $\tilde{y}_1$ with respect to x_i and x_j. In the first stage, the features obtained by the feature generator are also classified as true and false. In the method proposed in this section, a fully connected layer is used for classification and a binary cross entropy loss is used for supervision. The deentanglement pre-training process based on multi-supervision strategy proposed in this section is as follows:

Algorithm 1 Unentanglement pre-training based on multi-supervised strategy

---

Input: Face image *I*;

Output: Disentanglement feature *F*;

1: Through the output result *tildey*$_i$, the separated feature graph *F* (including real feature and attack feature) is obtained.

2: The feature graph *F* is passed through the fully connected layer to obtain feature *z*;

3: Computes the Hessian matrix of the feature generator *G* with respect to feature *x*;

4: Input feature *x* into the fully connected layer for binary classification;

5: In the other branch, the real features and attack features in the feature graph *F* are input

respectively to generate adversarial network branches for adversarial training:

6: Based on classified losses, Hessian penalises losses and generates counter-supervised loss back-passes, updating network parameters.

DIRE is effective in identifying images generated by diffusion models because diffusion models leave specific patterns in the alteration of pixel values when generating images. These patterns manifest as specific distributions of eigenvalues in the second derivatives of the Hessian matrix. By comparing the eigenvalues of the Hessian matrices of generated images and real images, DIRE can detect these discrepancies, thus distinguishing between images generated by diffusion models and real images.

3.3.3 Multiple classification stages based on DIRE features

Fine-grained classification often requires certain basic conditions, that is, it needs to have distinguishing features to ensure the accuracy and reliability of the model. Pre-training can provide good initial characteristics. The second stage of the two-stage coarse-to-fine multi-classification face anti-counterfeiting method proposed in this chapter requires the first stage to provide pre-trained feature generators to obtain the initial deentangled features for multi-classification. Furthermore, to emphasise the features of various types of attacks, this section strengthens the attack characteristics by multiplying an amplification factor λ, for the purpose of increasing the significance of the attack indicators in the artificially created facial samples. The enhancement of attack clues can further help the network to conduct multi-classification. Specifically, when a forged face I is input, the feature generator G trained in the first stage can get DIRE feature F_{1-4} to a certain extent.

$$F_{1-4} = G(I) \tag{17}$$

After the input face image passes through the feature generator, it is classified through softmax employing a completely integrated layer, and the output model predicts the probability of a certain category. The general classification task is to use the one with the highest probability as the prediction category and supervise the classification cross entropy loss. In fine-grained classification, such a design focuses only on the maximum probability and ignores the probability difference between the different categories, that is, the model does not care about the nuances of the learned features, which can lead to overfitting and poor generalisation performance. At the same time, in the multi-stage strategy, if only the maximum probability is concerned, the model can not reflect the optimisation of feature extraction ability.

The idea of Ranking loss is to compare the relative differences between two samples in order to learn the small differences between the samples. In this section, the supervision of Ranking loss can be understood as that in the second stage, the model's confidence in the prediction of true and false samples is higher than that in the first stage, so as to reflect the optimisation effect of the feature generator in the second stage. Specifically, when a forged face sample is input, suppose that the probability of the classification prediction as the real sample in the first stage is p, and the probability of the classification prediction as the real sample in the second stage is $\tilde{p} < p$, which needs to be met, which means that the probability of the model in the second stage predicting the forged face sample as the real face should be smaller, as shown in the following formula.

$$L_{Rank} = \max\left(\tilde{p} - p, 0\right) \tag{18}$$

Finally, as in the first stage, Hessian punishment is added in the second stage to supervise and ensure that the learned features are de-entangled.

Algorithm 2 Multiple classification stages based on DIRE features

Input: The first stage extracts the disentanglement feature F:

Output: Sample feature classification results;

The attack features in F are strengthened.

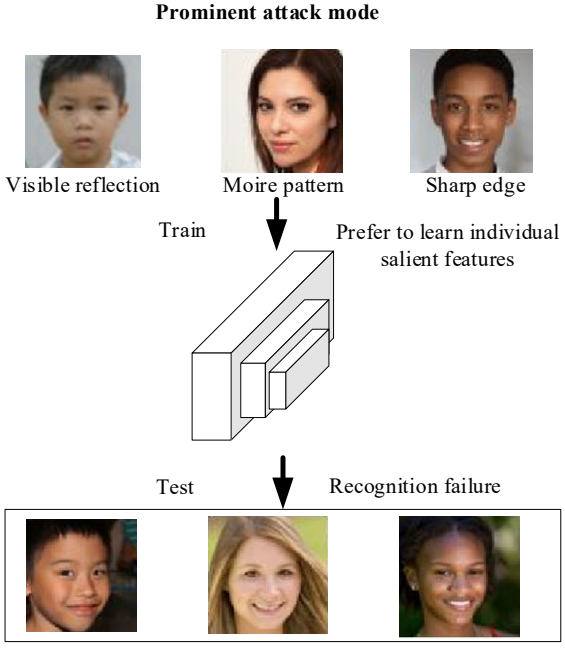
2: The forged face image / with enhanced attack clues is reconstructed by reconstructing network R;

Input / to the feature generator G to get F;

4: The real feature of F is extracted and the consistency constraint is applied to the first stage;

Input F into the multi-classification classifier,based on the multi-classification loss, Hessian penalises the loss and Ranking the loss backpass, and updates the network parameters.

Figure 4 The influence of salient features on model learning (see online version for colours)



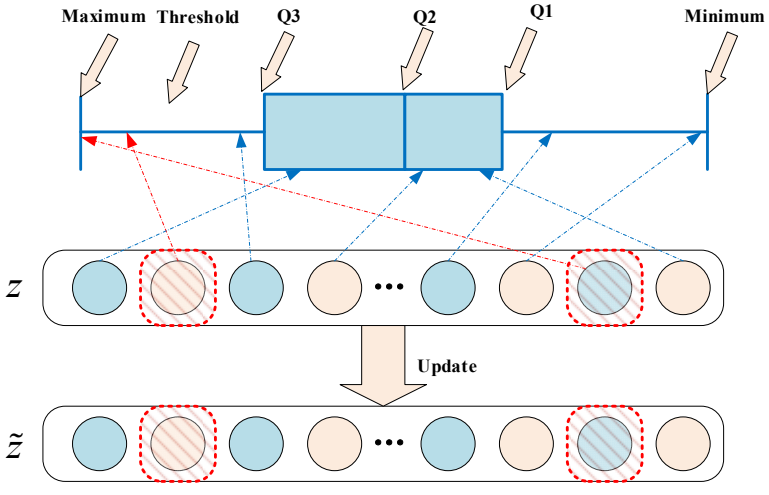
3.3.4 Abnormal gradient suppression algorithm

In the training, the model is only tested on the image data that obey the independent same distribution. But this type of testing often results in models with poor generalisation. In the face anti-counterfeiting task, this problem also needs to be considered, the domain deviation of the face anti-counterfeiting data set will also affect the learning of the network, different data domains will have special features, such as in some specific data domains, video replay attack data due to excessive superimposed reflection factors appear

strong highlights, There are also some data due to equipment factors appear obvious moire, photo printing attack processing should not appear obvious paper edge and so on. These features are prominent features that appear in a particular domain, but are not obvious in other domains. If unconstrained, the convolutional neural network will tend to focus on these prominent features in the training process, instead of learning general generalisation features, that is, falling into shortcut learning, which will cause the algorithm to fail when testing the universal attack pattern sample. As shown in Figure 4.

The abnormal gradient suppression training method in this section is designed to eliminate the influence of some prominent features in the optimisation process. This method belongs to an optimisation training method, which interferes during the convolutional neural network backpropagation to update parameters, calculates the backpropagation gradient during the first backpropagation, selects the network parameters corresponding to the abnormal gradient to cover up, and then updates the network parameters through the second forward propagation and back propagation. The aim is to improve the model's capacity to identify and apply various characteristics, avoid the model to overfit a few prominent features, and thus enhance the generalisation.

Figure 5 Diagram of updating vector z suppressed by anomaly gradient (see online version for colours)



Specifically, the input image gets domain independent feature X after the double normalisation module, and then gets a 512-dimensional feature vector z through the full connection layer. At the same time, a feature vector m of the same size is defined as a mask, which is used to cover up the abnormal network parameters selected later. The gradient of the loss with respect to z can then be obtained by formulas (3) to (14).

$$g_z = \frac{\partial L(z, y:\theta)}{\partial z} \tag{19}$$

where y is the second-class label of the face image, θ is the parameter of the network, and L is the objective function. In this instance, the training employs second-class differential entropy loss as the target function, as demonstrated in formula (3) to (15).

$$L = -\frac{1}{N}\sum_{n=1}^{N}\left[y^i \log\left(p_i\right) + \left(1-y^i\right)\log\left(1-p_i\right)\right] \tag{20}$$

In this section, a threshold is derived from adaptive calculation based on the quartile distance of the gradient data for filtering abnormal gradients. For this task, among all the outliers, the part that is too small has a slight effect on the update of network parameters, while the part that is too large has a greater impact on the update of network parameters. Consequently, focusing solely on the unusually high gradient value is essential. The process can be described as follows: Firstly, the gradient value of the loss with respect to z is calculated through formulas (3) to (14), and then the batch of gradient values are sorted, as shown in Figure 5.

The box plot is used to show an example of this batch of gradient distributions, where Q3, Q2, and Q1 represent the upper, middle, and lower quartiles, i.e., 75%, 50%, and 25% of the values corresponding to the gradient ordering. Then a threshold is calculated adaptively according to the gradient distribution, as shown in formulas (3) to (16).

$$Threshold = Q3 + \eta\left(Q3 - Q1\right) \tag{21}$$

where η is a non-negative weight parameter that adjusts the range of normal gradient values. Then the values with abnormally large gradients in the eigenvector z are covered. Firstly, the feature vector m is updated. For the part whose gradient is greater than the threshold value, the position i corresponding to the vector m is updated to 0, as shown in formulas (3) to (17).

$$m(i) = \begin{cases} 0, g_z(i) > Threshold \\ 1, Other \end{cases} \tag{22}$$

Then, the feature vector z and m are multiplied to get the updated vector $\tilde{z}$, thus achieving the purpose of concealing the prominent features.

$$\tilde{z} = z \odot m \tag{23}$$

where $\odot$ represents dot multiplication, the position of 0 in feature vector m is the value corresponding to abnormally large gradient, and the value at the same position of feature vector z needs to be covered. The parameters are not updated during the first backpropagation, but only used to update vector z. The updated ones will go through forward propagation again, and the network parameters will be updated during the second backpropagation. In addition, when all gradients are less than or equal to the threshold, that is, there is no abnormal gradient, and the vector z does not need to be updated. In other words, the abnormal gradient suppression proposed in this paper only works on prominent features.

4 Experimental analysis

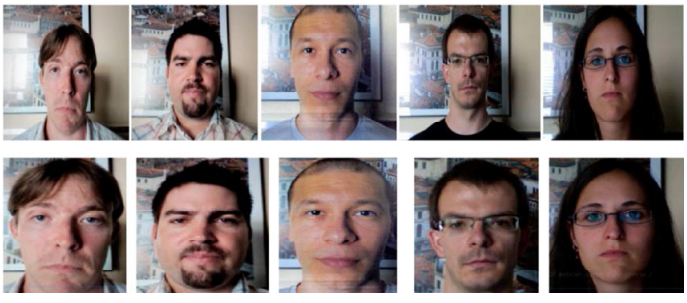
4.1 Data preprocessing and evaluation methods

In this section, the efficiency of the algorithm is assessed using four datasets accessible to the public in the realm of facial anti-counterfeiting,including OULU-NPU (denoted O),

CASIA FASD (denoted C), Idiap Replay-Attack (denoted I), and MSU-MFSD (denoted M). When conducting a single set of experiments, one data set is chosen for the target domain, while other data sets are designated as the source domain. Training of the model utilises the source domain data set, while the target domain data remains unseen. Then tests are conducted on the target domain data set to verify the generalisation performance of the model. As an illustration, datasets O, C, and I serve as the training source domain, while dataset M acts as the testing target domain. This set of experiments is recorded as O&C&I to M. Similarly, three other sets of experiments can be obtained, namely: O&C&M to I, O&M&I to C, and I&C&M to O.

Since the four original data sets are all video data, in order to obtain the input format of the model in this chapter, it is necessary to preprocess the video data and crop out the face picture. In this chapter, the Dlib package is used to process the original sample, and the output size of 256×256×3 face pictures is used as the input of the model, as shown in Figure 6. The first line represents the single frame image of the original video, and the second line represents the face picture captured by Dlib package. Attach a label to the obtained sample, the real face is 0, and the printed photo attack is 1.

Figure 6 Data preprocessin (see online version for colours)



4.2 Ablation experiment

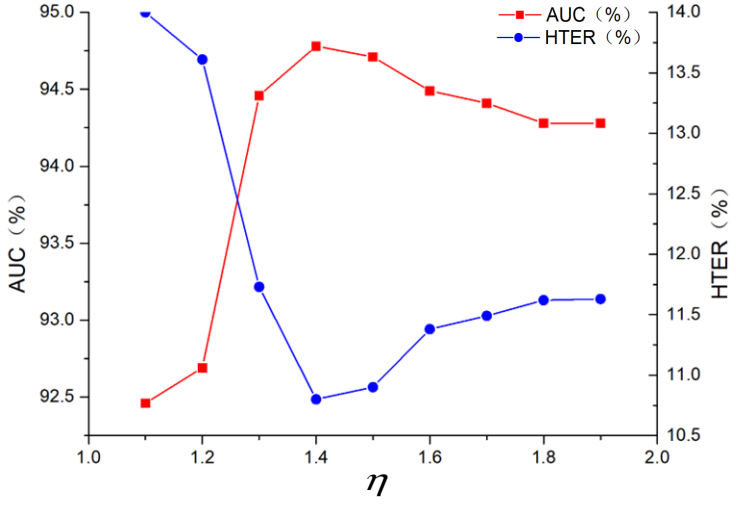
To verify the validity of each module of the method presented in this chapter, an ablation experiment was performed. Compare the effectiveness of each module and its impact on overall performance by cutting each module one by one. Baseline indicates the baseline model. Only the algorithm using Resnet18 as the backbone network is used to test the performance. Ours represents the DIRE face anti-counterfeiting method proposed in this chapter; Ours (w/o DNM) represents the method of unentanglement pre-training module without multi-supervision strategy proposed in this chapter, that is, only abnormal gradient suppression algorithm is added under the baseline model; Ours (w/o AGS) represents the method without anomaly gradient suppression proposed in this chapter, that is, the algorithm of the de-entanglement pre-training module with only more supervision strategies added under the baseline model. The results are shown in Table 2.

Table 2 Ablation experiment

Index	*Experiment*	*O&C&M to I*	*O&C&I to M*	*O&M&I to C*	*I&C&M to O*
HTER (%)	Baseline	24.46	17.57	29.38	23.37
	Ours (w/o DNM)	24.71	11.3	24.11	21.84
	Ours (w/o AGS)	22.43	11.67	18.45	22.4
	Ours	20.65	10.82	15.65	17.26
AUC (%)	Baseline	77.56	89.93	80.2	84.83
	Ours (w/o DNM)	75.03	93.5	85.53	86.1
	Ours (w/o AGS)	76.33	94.16	90.14	85.92
	Ours	81.1	94.88	92.66	89.93

The aim is to investigate how varying levels of gradient suppression affect the algorithm's efficiency, a search experiment was conducted on parameter η, which grew uniformly from 1.1 to 1.9 with step size of 0.1.

Figure 7 Effect of parameter η on AUC and HTER indicators (see online version for colours)



As can be seen from Figure 7, when η becomes larger from small to large, the performance decreases first and then increases. In formulas (3) to (16), when η is small, the threshold of abnormal gradient will be smaller, resulting in more gradients judged as abnormal, which will cover up more characteristic values and contain some useful information, resulting in performance degradation. When η gradually increases, the proportion of useful information in the covered features gradually decreases, and some features that are not conducive to generalisation performance are covered, so that the performance gradually increases to reach the optimal level. When η continues to increase, the inhibition effect on abnormal gradient gradually weakens, and the performance deteriorates again. This is because the threshold is too large, the judgment on abnormal gradient becomes more and more strict, and the inhibition effect decreases. When the threshold is greater than the maximum value of the ranking gradient, there will be no inhibition effect at all. When η is 1.4, both ACC and ACER are optimised, so η is finally set to 1.4 in this chapter.

Table 3 The comparison between the proposed method and existing methods on HTER (%) and AUC (%) evaluation criteria

Method	*O&C&M to I*		*O&C&I to M*		*O&M&I to C*		*I&C&M to O*	
	HTER (%)	*AUC (%)*	*HTER (%)*	*AUC (%)*	*HTER (%)*	*AUC (%)*	*HTER (%)*	*AUC (%)*
CNN	34.47	65.88	29.25	82.87	34.88	71.94	29.61	77.54
Auxiliary	29.14	71.67	22.72	85.88	32.53	73.16	30.18	77.62
MADDG	22.2	84.99	17.69	88.06	24.5	84.51	27.98	80.02
RFM	17.3	90.5	13.89	93.98	20.27	88.16	16.45	91.16
D2AM	15.43	91.22	12.7	95.66	20.98	85.58	15.27	90.87
NAS-FAS	14.5	93.8	19.53	88.63	16.54	90.16	13.8	93.42
Ours	20.65	81	10.83	94.86	15.64	92.67	17.28	89.95

4.3 Algorithm comparison

This paper compares several representative face anti-counterfeiting methods. In the tasks of O&C&I to M and O&M&I to C, the method proposed in this paper performs best.

This document compares a variety of typical facial anti-counterfeiting techniques. In the tasks of O&C&I to M and O&&I to C, the proposed method has the best performance, and the HTER index has increased by 1.88% and 5.34%, respectively. Compared with CNN, which uses manual features, and Auxiliary, an algorithm that uses depth map auxiliary, the algorithm introduced in this section clearly enhances the efficiency of the four tasks., which indicates that for cross-domain testing, the difference in the distribution of data in different domains has a great impact on the traditional manual feature method and the conventional depth model based method. The solution proposed in this chapter, which is designed for cross-domain problems, can effectively improve the generalisation performance of the algorithm. Some existing algorithms based on generative adversarial, domain adaptation and domain generalisation are also designed for cross-domain testing tasks. Compared with them, the method proposed in this chapter is also very competitive. Compared with MADDG and other algorithms based on adversarial learning that use domain tag assisted supervision, the performance of this algorithm is almost ahead of them in the four tasks, which also verifies the efficiency of the technique suggested in this study to extract generalised features without relying on domain tags. Compared with NAS-FAS, the proposed method has a considerable degree of competitiveness on the whole. On the one hand, among the four tasks, the proposed method performs better in O&C&I to M and O&M&I to C.

5 Results and discussion

This document concentrates on developing a universal detector for differentiating facial images produced through diffusion. Our findings indicate that earlier created image detectors had restricted efficacy in identifying images produced by diffusion models. Addressing this issue, we suggest a method for image depiction that relies on the inaccuracies in reconstructing DDIM inverted and reconstructed images. At the same time, the original features of the face are constrained to eliminate the domain features from the features of different source domains, standardise the feature space, and initially enhance the generalisation performance of the model. At the same time, an abnormal gradient suppression algorithm is proposed in the process of model training, which adaptively calculates filtering gradients in the process of parameter optimisation, detects abnormally large gradient values, and supposes them to avoid shortcut learning by the network and instead learns more generalised features instead of prominent features. Improve the model's ability for broader generalisation. Our aspiration is that our research will establish a robust foundation for detecting images produced by diffusion. Four data sets were selected in this paper, one as the target domain and the other three as the source domain for cross-domain testing. The four groups of experiments were recorded as O&C&M to I, O&C&I to M, O&&ITO C and I&C&M to O. Firstly, various parts of the method proposed in this chapter are added and subtraction, and ablation experiments are conducted to verify its effectiveness. Then, the algorithm proposed in this chapter is compared with several representative face anti-counterfeiting methods. The results show that the algorithm proposed in this chapter performs best on O&C&I to M and O&M&I

to C, and has great competitive advantages on the whole. Concurrently, the intricate nature of the comparative experiment indicates that the technique discussed in this section offers a more substantial computational benefit.

Table 4 Comparison between the proposed method and the existing methods in terms of parameters and calculation amount

Methods	*Argument* ($\times 10^6$)	*Floating point arithmetic* ($\times 10^9$)
Auxiliary	2.22	93.14
STDN	1.3	80.1
D2AM	1.33	26.38
RFM	3.9	47.91
NAS-FAS	2.58	53.66
Ours	5.66	22.78

All in all, the method proposed in this paper is a competitive and generalised AI-generated face image verification method in unknown scenes

Acknowledgements

This work was supported by National Natural Science Foundation of China (grant no. 42202125), Natural Science Foundation of Sichuan Province (grant no. 2024NSFSC0828).

Declarations

The author declares that it does not have any known interests or personal relationships that could potentially influence the reported work in this paper.

References

Bengio, Y., Courville, A.C. and Vincent, P. (2012) 'Representation learning: a review and new perspectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August, Vol. 35, No. 8, pp.1798–1828.

Bharadwaj, S., Dhamecha, T.I., Vatsa, M. and Singh, R. (2013) 'Computationally efficient face spoofing detection with motion magnification', *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.105–110.

Bhatt, P., Sethi, A., Tasgaonkar, V., Shroff, J., Pendharkar, I., Desai, A., Sinha, P., Deshpande, A., Joshi, G., Rahate, A., Jain, P., Walambe, R., Kotecha, K.V. and Jain, N.K. (2023) 'Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions', *Brain Informatics*, 31 July, Vol. 10, No. 1, p.18.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D. and Wallace, E. (2023) 'Extracting training data from diffusion models', *ArXiv, abs/2301.13188.*

Chen, T.Q., Li, X., Grosse, R.B. and Duvenaud, D.K. (2018) *Isolating Sources of Disentanglement in VAEs*, MIT Press, Montreal.

Chen, W., Hu, H., Saharia, C. and Cohen, W.W. (2022) 'Re-imagen: retrieval-augmented text-to-image generator', *ArXiv, abs/2209.14491.*

Dhariwal, P. and Nichol, A. (2021) 'Diffusion models beat GANs on image synthesis', *ArXiv, abs/2105.05233.*

Esser, P., Rombach, R. and Ommer, B. (2020) 'Taming transformers for high-resolution image synthesis', *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp.12868–12878.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) 'Generative adversarial nets', *Advances in Neural Information Processing Systems*, Vol. 27, p.27.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A.C. (2017) 'Improved training of Wasserstein GANs', *Neural Information Processing Systems*, Vol. 31.

He, T. and Qin, F. (2024) 'Exploring how the metaverse of cultural heritage (MCH) influences users' intentions to experience offline: a two-stage SEM-ANN analysis', *Heritage Science*, Vol. 12, pp.1–18.

Henzler, P., Mitra, N.J. and Ritschel, T. (2018) 'Escaping Plato's cave: 3D shape from adversarial rendering', *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.9983–9992.

Ho, J., Jain, A. and Abbeel, P. (2020) 'Denoising diffusion probabilistic models', ArXiv, *abs/2006.11239*.

Huang, H., Sun, L., Du, B., Fu, Y. and Lv, W. (2022) 'GraphGDP: generative diffusion processes for permutation invariant graph generation', *2022 IEEE International Conference on Data Mining (ICDM)*, pp.201–210.

Liu, L., Ren, Y., Lin, Z. and Zhao, Z. (2022) 'Pseudo numerical methods for diffusion models on manifolds', *ArXiv, abs/2202.09778.*

Liu, S. and Deng, W. (2015) 'Very deep convolutional neural network based image classification using small training sample size', *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp.730–734.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Scholkopf, B. and Bachem, O. (2018) 'Challenging common assumptions in the unsupervised learning of disentangled representations', *International Conference on Machine Learning.*

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C. and Zhu, J. (2022) 'DPM-Solver++: fast solver for guided sampling of diffusion probabilistic models', *ArXiv, abs/2211.01095.*

Mo, H., Chen, B. and Luo, W. (2018) 'Fake faces identification via convolutional neural network', *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security.*

Nichol, A. and Dhariwal, P. (2021) 'Improved denoising diffusion probabilistic models', *ArXiv, abs/2102.09672.*

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I. and Chen, M. (2021) 'GLIDE: towards photorealistic image generation and editing with text-guided diffusion models', *International Conference on Machine Learning.*

Ricker, J., Damm, S., Holz, T. and Fischer, A. (2022) 'Towards the detection of diffusion model deepfakes', *ArXiv, abs/2210.14571.*

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2021) 'High-resolution image synthesis with latent diffusion models', *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.10674–10685.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M. and Aberman, K. (2022) 'DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation', *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.22500–22510.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J. and Norouzi, M. (2022) 'Photorealistic text-to-image diffusion models with deep language understanding', *ArXiv, abs/2205.11487.*

Shaul, N., Singer, U., Chen, R.T., Le, M., Thabet, A.K., Pumarola, A. and Lipman, Y. (2024) 'Bespoke non-stationary solvers for fast sampling of diffusion and flow models', *ArXiv, abs/2403.01329.*

Shiohara, K. and Yamasaki, T. (2022) 'Detecting deepfakes with self-blended images', *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.18699–18708.

Sohl-Dickstein, J.N., Weiss, E.A., Maheswaranathan, N. and Ganguli, S. (2015) 'Deep unsupervised learning using nonequilibrium thermodynamics', *ArXiv, abs/1503.03585.*

Song, J., Meng, C. and Ermon, S. (2020) 'Denoising diffusion implicit models', *ArXiv, abs/2010.02502.*

Song, Y., Dhariwal, P., Chen, M. and Sutskever, I. (2023) 'Consistency models', *International Conference on Machine Learning.*

Stehouwer, J., Dang, H., Liu, F., Liu, X. and Jain, A.K. (2019) 'On the detection of digital face manipulation', *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5780–5789.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2015) 'Rethinking the inception architecture for computer vision',*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2818–2826.

Tashiro, Y., Song, J., Song, Y. and Ermon, S. (2021) 'CSDI: conditional score-based diffusion models for probabilistic time series imputation', *Neural Information Processing Systems.*

Wang, S., Wang, O., Owens, A., Zhang, R. and Efros, A.A. (2019a) 'Detecting photoshopped faces by scripting photoshop', *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.10071–10080.

Wang, S., Wang, O., Zhang, R., Owens, A. and Efros, A.A. (2019b) 'CNN-generated images are surprisingly easy to spot for now', *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.8692–8701.

Xiong, W., Yu, C., Shi, C., Zheng, Y., Wang, X., Hu, Y., Yin, H., Li, C. and Wang, C. (2024) 'V4RIN: visual analysis of regional industry network with domain knowledge', *Visual Computing for Industry, Biomedicine, and Art*, Vol. 7, Article number: 11.

Xu, Y., Deng, M., Cheng, X., Tian, Y., Liu, Z. and Jaakkola, T. (2023) 'Restart sampling for improving generative processes', *ArXiv, abs/2306.14878.*

Yu, F., Zhang, Y., Song, S., Seff, A. and Xiao, J. (2015) 'LSUN: construction of a large-scale image dataset using deep learning with humans in the loop', *ArXiv, abs/1506.03365.*

Zhu, D., Chen, D., Grossklags, J. and Fritz, M. (2023) 'Data forensics in diffusion models: a systematic analysis of membership privacy', *ArXiv, abs/2302.07801.*