ITC 2/53
Information Technology and Control
Vol. 53 / No. 2 / 2024
pp. 408-428
DOI 10.5755/j01.itc.53.2.35590

GAN-Generated Face Detection Based on Multiple Attention Mechanism and Relational Embedding

Received 2023/11/11 | Accepted after revision 2024/03/02

HOW TO CITE: Ouyang, J., Ma, J., Chen, B. (2024). GAN-Generated Face Detection Based on Multiple Attention Mechanism and Relational Embedding. *Information Technology and Control,* 53(2), 408-428. https://doi.org/10.5755/j01.itc.53.2.35590

GAN-Generated Face Detection Based on Multiple Attention Mechanism and Relational Embedding

Junlin Ouyang
School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China;
Hunan Key Laboratory for Service computing and Novel Software Technology, Xiangtan, China;
Hunan Software Vocational and Technical University, Xiangtan, China; e-mail: yangjunlin0732@163.com

Jiayong Ma
School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China;
Hunan Key Laboratory for Service computing and Novel Software Technology, Xiangtan, China

Beijing Chen
School of Computer and Software, Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, China

Corresponding author: yangjunlin0732@163.com

The rapid development of the Generative Adversarial Network (GAN) makes generated face images more and more visually indistinguishable, and the detection performance of previous methods will degrade seriously when the testing samples are out-of-sample datasets or have been post-processed. To address the above problems, we propose a new relational embedding network based on "what to observe" and "where to attend" from a relational perspective for the task of generated face detection. In addition, we designed two attention modules to effectively utilize global and local features. Specifically, the dual-self attention module selectively enhances the representation of local features through both image space and channel dimensions. The cross-correlation attention module computes similarity between images to capture the global information of the output in the image. We conducted extensive experiments to validate our method, and the proposed algorithm can effectively extract the correlations between features and achieve satisfactory generalization and robustness in generating

face detection. In addition, we also explored the design of the model structure and the inspection performance on more categories of generated images (not limited to faces). The results show that RENet also has good detection performance on datasets other than faces.

KEYWORDS: Image forensics; forgery detection; GAN-generated face detection; generative adversarial networks; relational networks.

1. Introduction

The Generative Adversarial Network (GAN) [18] has been gradually applied to many fields since it was proposed [34, 39, 57]. With its rapid development, the generated images are becoming more and more realistic, as shown in Figure 1 for the face images generated by GAN. Such GAN-generated face images are difficult to distinguish from human eyes and can be easily generated by ordinary people. If they are used for malicious purposes, it may adversely affect some individuals' reputations and even social security ethics. Therefore, the detection of GAN-generated face images has become increasingly necessary.

In recent years, some researchers [27, 50, 51] have verified the authenticity of images by actively embed watermarks to the images. Besides, many passive forensic methods have been proposed. In [1, 7, 16, 19, 23, 45, 49, 59] they detect natural and generated faces by exploring the differences in the image formation process. They combine traditional forensic methods to generated face detection. However, when confronted with fake face images where only local areas are generated, searching for feature differences directly on the entire generated face image may lead to detection failure. Therefore, [4, 5, 8, 38] also combine local information such as artifacts to assist in detection. Although the methods described above achieve relatively high detection accuracy, they suffer from poor generalization and a lack of interpretability [21]. [21, 22, 26, 64] try to make the results interpretable by looking for inconsistencies in the physiology-based methods. However, with the continuous innovation of GAN, the difference between generated images and natural images in the spatial domain becomes increasingly difficult to detect [68]. As a consequence, [6, 9, 13, 14, 17, 37, 43, 68] turned their attention to the frequency domain, which improves the detection generalization performance by fusing features from the spatial and frequency domains. But their methods cannot adaptively capture the most dis-criminative features.

To address the above challenges, we propose a method that is inspired by relational networks [56] and combines both dual-self attention (DSA) and cross-correlation attention (CCA) to learn "what to observe" and "where to attend" from an image relations perspective. We achieve this goal by utilizing relational patterns within and between images through the Relational Embedding Network (RENet).

The DSA module learns its own feature associations for the purpose of enriching semantic in-formation. The CCA module calculates the correlations between images so that they can have global information. Our contribution can be summarized as follows:

1. We propose a Relation Embedding Network (RENet) with a multi-attention mechanism to improve the ability of detecting GAN-generated face.
2. We design dual-self attention and cross-correlation attention to enhance the local spatial and channel-wise correlated features within an image, and to link global relationships between images, respectively.
3. We conduct extensive experiments to verify the proposed method has excellent generalization ability and robustness to common post-processing operations.

Figure 1

Some samples in the experimental datasets. From left to right, the columns are the fake faces generated by ProGAN [30], StyleGAN [31], StyleGAN2 [32], StarGAN [11], BeGAN [2], LsGAN [46], WgGANGP [20], RelGAN [42] respectively. The generated faces are indistinguishable by human eyes

2. Related Work

Fake face detection. With the development of GAN technology and fake face technology, researchers have proposed many methods to detect fake faces. Hu et al. [26] and Guo et al. [21] found that the highlight of the two eyes in GAN-generated faces are inconsistent. Guo et al. [22] found that the pupil shape of real faces should be close to circular or elliptical, while the pupils of generated faces show irregular and inconsistent shapes. Although physiology-based inconsistency detection is interpretable, accuracy is greatly reduced if the inconsistencies are occluded or the scene angle is biased. Nataraj et al. [49] extracted the co-occurrence matrices on the RGB channels of the image and input them into a neural network for classification. Barni et al. [1] reported a significant impact of the correlation between color channels on detection effectiveness. Besides individual RGB channels, they also calculated the co-occurrence matrices from pairwise combinations of channels. The experimental results showed that multiple channels can further improve the robustness of detection compared to using single color channel information. However, the accuracy and robustness of detection only in the RGB domain are far from satisfactory. Chen et al. [7] combined dual color spaces and designed an improved Xception network model to increase detection robustness. Guo et al. [23] adaptive convolution to predict manipulation traces in an image, and then maximized manipulation artifacts by updating weights through backpropagation. Liu et al. [45] analyzed that it was more robust to detect fake faces by texture, so they proposed GramNet to capture long-range texture information to improve the robustness and generalization of the model. Despite the commendable detection performance exhibited by previous methods, relying solely on the spatial domain still presents limitations [14]. Moreover, the frequency domain of images has been widely applied in various fields [10, 54, 69, 71, 72]. Frank et al. [14] found that there are significant differences in the DCT spectra of real and fake images, and the DCT spectrum is more robust for detecting image manipulation than the RGB spectrum. Liu et al. [43] argued that upsampling is a necessary step in most face-forgery techniques, which leads to significant changes in the frequency domain, particularly the phase spectrum. Thus, they captured upsampling artifacts in face-forgery by combining spatial images with phase spectra, achieving a good detection result. Luo et al. [67] discovered that noise in face regions is continuously distributed in real images, while in manipulated images, it appears smoother or sharper. Therefore, they employed the high-pass filter SRM to extract high-frequency noise for detecting face forgery. Le et al. [35] utilized frequency-domain knowledge distillation to retrieve the removed high-frequency components in the student network for enhancing the detection accuracy of low-resolution images. Although analyzing in other domains can improve the accuracy and robustness of detection, they focus on the global features of the image and are typically difficult to detect subtle local tampering. Chai et al. [4] proposed the patchCNN network, which truncates the entire network to focus on local artifacts. Experiments have shown that local texture information can enhance the model's generalization ability. Jia et al. [28] designed a dual-branch network for predicting image-level and pixel-level fake labels based on inter-image and intra-image inconsistency, which is processed by stable wavelet decomposition. Ju et al. [5] introduced the FPN modules into Xception and reduced the number of convolutional layers in the network to detect locally generated face. The model has good performance for detecting faces with small generated regions. However, they overemphasized local features and ignored the relationship between global and local features. In contrast to the above, we integrate both global and local features, making the framework more robust and demonstrating better generalization ability.

Attention model. The human visual nerve receives more data than it can process, so it requires the human brain to weigh the inputs and focus only on the necessary information [24]. For this reason, the researchers used a similar concept in their experiments. Vaswani et al. [58] first proposed a self-attention mechanism and applied it to machine translation, which reveals concerns about image structure through similarities within a domain. Recent work [40, 41, 48, 52, 62] has shown that the self-attention mechanism can effectively capture contextual relationships and improve the intra-class compactness of images. In addition to focusing on the connections between images, the relationships between images have been used to form a central part of various prob-

lems in computer vision. It calculates the relationship between two images and applies to video action recognition, few-shot classification, semantic segmentation, medical image segmentation, style transfer, etc. Recently, some GAN-generated detection algorithms [5, 7] have adopted self-attention mechanisms to enhance semantic information, but they have not accounted for connections between images. Inspired by Wu et al. [63], we introduce not only self-attention in each image but also cross-attention between images to enhance the ability to classify relevant regions.

Relation Network. The relational network [56] is a metric-based network structure with the core idea of mapping images to a learnable embedding function to extract features of interest and then distinguishing different classes by measuring the similarity of features between samples. Thus, attention can be used to correct and strengthen the feature regions of interest. In this paper, multiple attention mechanisms that are employed post the embedding function to rectify the network's focus on relevant areas and boost the expression of pertinent regions This approach promotes the network's generalization and robustness.

3. Approach

In this section, we will introduce the Relational Embedding Network (RENet), which is used to solve the problem of poor generalization and robustness in GAN-generated face detection. In Figure 2, the overall architecture consists of three modules: a basic representation module, a feature augmentation module, and a representation comparison module. The network parameters of feature augmentation module, and a representation comparison module are shown in Table 2 The basic representation module and the representation comparison module are similar to the modules of the relational network. On this basis, we add a feature enhancement module, which is composed of dual-self attention (DSA) and cross-correlation attention (CCA). We provide a brief overview of the overall RENet in Section 3.1. Then, we introduce the implementation details of DSA and CCA in Section 3.2 and Sec. 3.3, respectively.

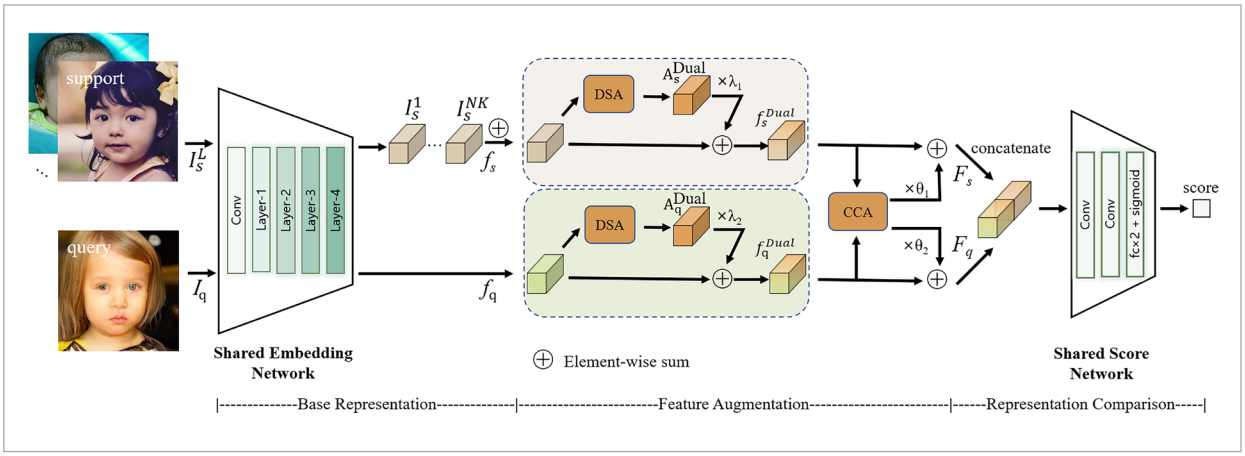
3.1. Architecture Overview

In this paper, we treat real images from the same dataset or fake images generated by same type of GAN as one domain, which is similar to the setting of other related tasks [19, 58]. The training set for each domain is denoted D_{train}, and the test set is denoted D_{test}. Both D_{train} and D_{test} are divided into multiple episodes for training, each of them contains a query set $Q = (I_q, y_q)$ and a support set $S = \{(I_S^L, y_S^L)\}_{L=1}^{NK}$ with N categories and K images per category.

As shown in Figure 2, given a pair of support and query set images $\{I_S^L, I_q\}$, each of them has a size of C×H×W. They pass through a shared embedding network and generate corresponding features $f_s^1, ..., f_S^{NK}$ and f_q.The

Figure 2

Overall architecture of RENet. We feed f_S and f_q from shared embedding network into the DSA module to obtain locally enhanced features f_s^{Dual} and f_q^{Dual}. Following CCA, we can obtain F_S and F_q, both of which contain global information. Finally, they are concatenated along the channel and the category with the highest score determines the classification result



support set features f_s^1, ..., f_S^{NK} from the same domain will be ∈ denoted as f_s via element-wise sum. The DSA module first applies self-attention over them to generate self-attentive features $\{A_s^{Dual}, A_q^{Dual}\} \in \mathbb{R}^{C\times H\times W}$. and multiplies them by their corresponding weights before adding them to the input features to obtain $f_{s\text{-}Dual}$ and f_q^{Dual}, respectively. The resulting features are then processed by the CCA module to generate $\{A_s^{Cross}, A_q^{Cross}) \in \mathbb{R}^{C\times H\times W}$, which operates similarly to the DSA module. From there, output feature F_s and F_q are computed for the classification score by representation comparison, with F_s being calculated as follows:

$$\begin{cases} f_s^{Dual} = f_s + \lambda_1 \times A_s^{Dual} \\ F_s = f_s^{Dual} + \theta_1 \times A_s^{Cross} \end{cases} \quad (1)$$

where λ_1 and θ_1 are learnable parameters for assigning weights, which gradually learn a weight from 0. The output features have their own local enhancements as well as correlation information between images. F_q is calculated in a similar way to F_s.

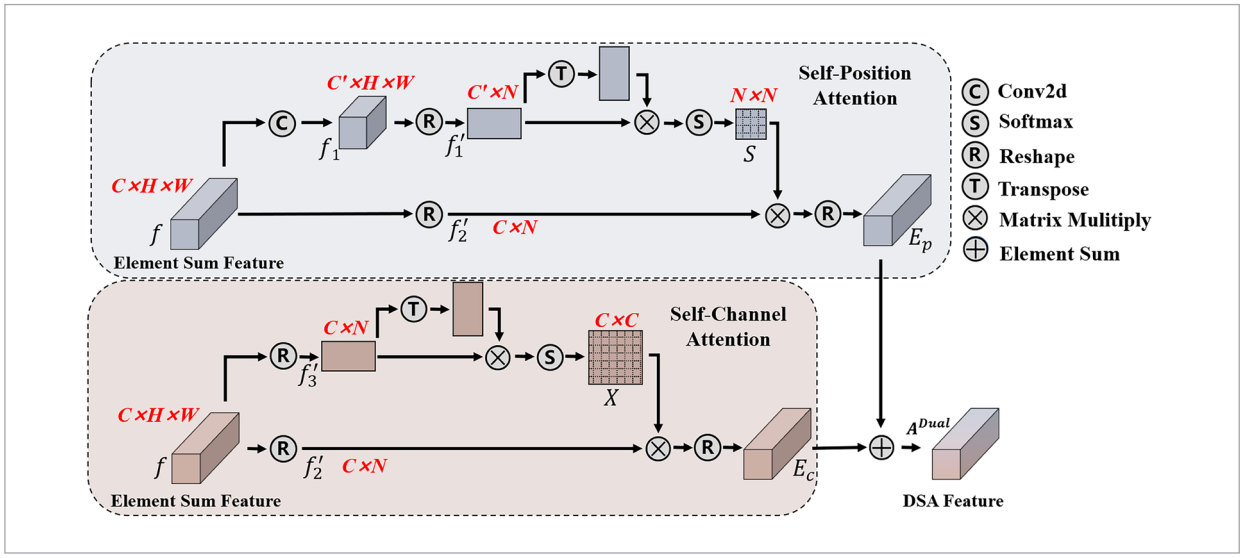
3.2. Dual-Self Attention (DSA)

Self-Position Attention(SPA). Global texture features help to detect GAN-generated faces and can improve model detection ability by capturing long-range texture information [15]. However, [16, 17] show that generated face detection relies more on subtle details in the hair or background when performing cross dataset detection, so focusing on local detailed features can effectively enhance the generalization ability of detection. In order to pay more attention to local detailed features, we introduce the self-position attention module, which lets local features establish links in contextual relationships, thereby enhancing the network's expression of local features.

As shown in the upper part of Figure 3, given the feature $f \in \mathbb{R}^{C\times H\times W}$, it is first fed into a 1x1 convolutional layer to generate $f_1 \in \mathbb{R}^{C'\times H\times W}$ where $(C'<C)$. We then reshape f_1 and f into $f_1' \in \mathbb{R}^{C'\times N}$ and $f_2' \in \mathbb{R}^{C\times N}$, where $N = H \times W$ in the SPA module. Next, f_1' is transposed to $f_1'^T \in \mathbb{R}^{N\times C'}$, and the matrix multiplication is performed between f_1' and $f_1'^T$ to obtain their relationship, denoted by $g(f_1', f_1'^T)$. After softmax layer, we will obtain position attention maps $S \in \mathbb{R}^{N\times N}$. From the spatial position, i of f_1' and j of $f_1'^T$, we respectively get two spatial points $\{f_{1i}', f_{1j}'^T\}$ $\mathbb{R}^{C'}$,where $i \in \{1, ..., N\}$, $j \in \{1, ..., N\}$. We further denote the pointwise calculation of $g(f_1', f_1'^T)$ as $g_{ij}(f_{1i}', f_{1j}'^T)$, and have the following equation:

Figure 3

The details of DSA. Features are fed to the self-position attention (SPA, upper) and self-channel attention (SCA, below) modules, which output E_p and E_c, respectively. They then fuse into features A_{Dual} with positional attention and channel attention. DSA aims to make the network better understand "what to observe"



$$s_{ji} = \frac{\exp\left(g_{ij}\left(f'_{1i}, f'^{T}_{1j}\right)\right)}{\sum_{i=1}^{N}\exp\left(g_{ij}\left(f'_{1i}, f'^{T}_{1j}\right)\right)} = \frac{\exp\left(f'_{1i}\cdot f'^{T}_{1j}\right)}{\sum_{i=1}^{N}\exp\left(f'_{1i}\cdot f'^{T}_{1j}\right)} \quad (2)$$

where $s_{ij} \in \mathbb{R}^{1\times1}$ represents feature at position i impact on position j. The more correlated the features at the two positions are, the larger s_{ij} is. Then, after multiplying with the matrix $f'_{2i} \in \mathbb{R}^{1\times N}$, we sum them to obtain the correlation of all features in column i, $E_j \in \mathbb{R}^{1\times N}$:

$$E_j = \sum_{i=1}^{N}(s_{ji}\cdot f'_{2i}). \quad (3)$$

Finally, we aggregate the features of all points and reshape to get the final output $E_p \in \mathbb{R}^{C\times H\times W}$, i.e.:

$$E_p = \sum_{j=1}^{N} E_j \quad (4)$$

where E_p selectively aggregates the context based on the position attention map. Similar semantic features are correlated with each other, thereby improving intra-class compactness and semantic consistency.

Self-Channel Attention (SCA). Each channel of high-level semantic features can be considered a particular class of response, and the different semantic responses are interrelated [15]. As a consequence, we add a self-channel attention module to build dependencies between channels.

As shown in the lower half of Figure 3, unlike the SPA module, before computing the relationship between the two channels, we directly reshape the input features to obtain $f_3' \in \mathbb{R}^{C\times N}$ and $f_2' \in \mathbb{R}^{C\times N}$ instead of feeding them into the convolutional layer, as this maintains the relationship between different channel mappings. f_3' is multiplied by its transposed feature map to produce channel attention feature map $X \in \mathbb{R}^{C\times C}$. The subsequent calculation is similar to equations (2), (3), and (4), the final output $E_C \in \mathbb{R}^{C\times H\times W}$ can be obtained by matrix multiplication of X and f_2', i.e.:

$$E_C = \sum_{j=1}^{N}\sum_{i=1}^{N}(x_{ji}\cdot f'_{2i}). \quad (5)$$

Finally, in order to fully fuse the local detail features between position and channel, we sum the features from these two attention modules to obtain fusion feature A^{Dual}. The DSA module does not add too many parameters to the network but effectively enhances the local representation of the features, allowing the network to better understand "what to observe."

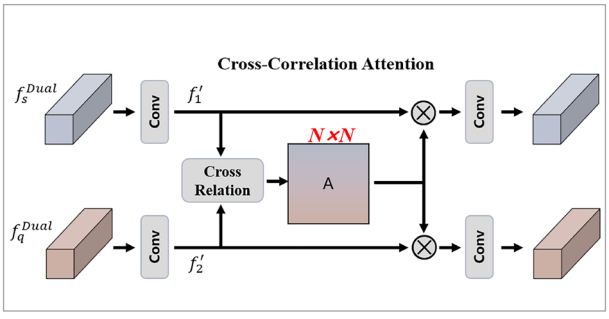
3.3. Cross-Correlation Attention (CCA)

The GAN model concentrates on learning local features and uses an upsampling process to produce visually enhanced images. However, this local-to-global structure deviates fundamentally from the global generation pattern of natural images, resulting in GAN-generated images frequently lacking global features. Such images may appear indistinguishable to the human eye, but in reality, they are merely a composite of uncorrelated local features.

After applying the DSA module, a pair of features that aggregate local information can be obtained. However, since there is no correlation between these two features, neglecting global information could potentially impact the detection results. Consequently, we integrated the CCA into the pipeline after the DSA module to establish inter-image links, facilitating detection and classification.

Figure 4

The details of CCA. Aim to adjust the "focus" of the image given during the network test



As shown in Figure 4, the features f_s^{Dual} and f_q^{Dual} are fed into a 1x1 convolution network to adjust the channel size. The output is then reshaped such that $f_1' \in \mathbb{R}^{C'\times N_1}$ and $f_2' \in \mathbb{R}^{C\times N_2}$,where $C'< C$, $N_1 = H_1\times W_1$, $N_2 = H_2\times W_2$ (Note, in this paper, $N_1 = N_2$. Here we write separately to better illustrate the processes involved). Denote the relationship between f_1' and f_2' as $g(f_1', f_2')$, we can compute their cross-attention graph maps $\boldsymbol{A}\in\mathbb{R}^{N_1\times N_2}$. Similar as SPA module, i of f_1' and j of f_2', we respectively get two spatial points $\{f'_{1i}, f'_{2j}\} \in \mathbb{R}^{C'}$, where $i \in \{1, ..., N_1\}$, $j \in \{1, ..., N_2\}$ and denote the pointwise calculation of $g(f_1', f_2')$ as $g_{ij}(f'_{1i}, f'_{2j})$. We choose the co-

sine similarity function to calculate the relationship between features:

$$A_{ij} = g_{ij}(f'_{1i}, f'_{2j}) = sim(f'_{1i}, f'_{2j}), \quad (6)$$

where sim(·,·) means the cosine similarity between two features and $sim(f'_{1i}, f'_{2j}) = \frac{f'_{1i} \cdot f'_{2j}}{\|f'_{1i}\| \cdot \|f'_{2j}\|}$.We perform l2-normalization over f_1' and f_2', along their channel dimension, then equation (6) can be rewritten as:

$$A = g(f'_1, f'_2) = sim(f'_1, f'_2). \quad (7)$$

The feature map contains the cross-correlation between each position in A and B. After obtaining the cross-feature map, it is multiplied by the corresponding matrices of f_1' and f_2',respectively:

$$\begin{cases} A_1^{Cross} = f'_1 \cdot \boldsymbol{A} \\ A_2^{Cross} = f'_2 \cdot \boldsymbol{A} \end{cases}. \quad (8)$$

It can be seen that the output feature A_1^{Cross} contains global information of f_1 for each pixel, and A_2^{Cross} is the same. CCA promotes the generation of more discriminative features for semantically similar regions between support and query images, allowing the network to adjust its "focus" on the images during testing. Finally, the channels of the features are adjusted to output features $A_1^{Cross} \in \mathbb{R}^{C \times H_1 \times W_1}$ and $A_2^{Cross} \in \mathbb{R}^{C \times H_2 \times W_2}$.

4. Experiments

In this section, we first introduce the details of the dataset and experiments. Then we conduct a series of contrast experiments and ablation experiments to analyze the effectiveness of the pro-posed model and modules. Finally, we discuss how to design the RENet model structure to achieve the best performance of the network. Additionally, we explore the detection effectiveness across various generated image categories, including food, animals, landscapes, and more.

4.1. Datasets

In the experiment, the real faces CelebA-HQ [44] and FFHQ [31] are used as positive samples, and the fake faces generated by PGGAN [30] and StyleGAN [31] are used as negative samples, i.e. the number of categories N = 4. Following [4], we randomly select 10K images from each of the four datasets and then divide the images into training, validation, and testing sets in a ratio of 7:1:2. Each image is resized to 256 × 256. Data augmentation proves to be effective in mitigating the overfitting problem in deep learning models and enhancing their generalization ability [60, 70]. Consequently, during training, we exclusively augment query images to simulate real-world scenarios.

To assess generalization ability, we randomly select 2,000 images from commonly used out-of-sample datasets as test sets including StyleGAN2 [33], StarGAN [11], BeGAN [2], LsGAN [46], WgGANGP [20], RelGAN [42] The details of the corresponding generated faces are provided in Table 1 and each image resized to 256 × 256.

Table 1
Details of datasets

Source	Models	Resolution
CelebA	StarGAN	256x256
	BeGAN	128x128
	LsGAN	128x128
	WgGANGP	128x128
CelebA-HQ	ProGAN	1024x1024
	RelGAN	256x256
FFHQ	StyleGAN	256x256
	StyleGAN2	1024x1024

4.2. Implementation Details

Network architectures. The complete network is partitioned into three segments: Base Representation, Feature Augmentation, and Representation Comparison. In the Base Representation, we discard the final average pooling layer and fully connected layer of ResNet50 [25] and use the remaining layers as the shared embedding network in the RENet. The structures and parameters of the remaining two segments will be thoroughly outlined in Table 2. In Figure 2, image $f \in \mathbb{R}^{3 \times 256 \times 256}$ is input into the shared embedding network to obtain the feature $f \in \mathbb{R}^{2048 \times 8 \times 8}$. In the score network, the features are first fed into two identical convolutional groups. Each group con-

Table 2
Network parameters of feature augmentation module and representation comparison

	Name	Parameters
Feature Augmentation	AdaptiveAvgPool2d_1	output_size=8
	Conv2d_1	out_channels=256, kernel_size=1, stride=1
	Conv2d_2	out_channels=256, kernel_size=1, stride=1
	Matrix_Mulitiply_1	(Conv2d_1,Conv2d_2)
	Softmax_1	softmax(Matrix_Mulitiply_1)
	Conv2d_3	out_channels =2048, kernel_size=1, stride=1
	Matrix_Mulitiply_2	(Softmax_1, Conv2d_3)
	Matrix_Mulitiply_3	(AdaptiveAvgPool2d_1, AdaptiveAvgPool2d_1)
	Softmax_2	softmax(Matrix_Mulitiply_3)
	Matrix_Mulitiply_4	(Softmax_2, Matrix_Mulitiply_3)
	Add_S, Add_Q	(Matrix_Mulitiply_2, Matrix_Mulitiply_4)
	Conv2d_4_1, Conv2d_4_2	out_channels =256, kernel_size=1, stride=1
	Conv2d_5_1, Conv2d_5_2	out_channels =256, kernel_size=1, stride=1
	Matrix_Mulitiply_5_1, Matrix_Mulitiply_5_2	(Conv2d_4_2, Conv2d_5_1) (Conv2d_4_1, Conv2d_5_2)
	Softmax_3_1, Softmax_3_2	Softmax_3_1(Matrix_Mulitiply_5_1), Softmax_3_2(Matrix_Mulitiply_5_2)
	Matrix_Mulitiply_6_1, Matrix_Mulitiply_6_2	(Softmax_3_1,Add_S), (Softmax_3_2,Add_Q)
Representation Comparison	Conv2d_6	out_channels =64, kernel_size=3, stride=1, padding=1
	MaxPool2d_1	kernel_size=2, stride=2
	Conv2d_7	out_channels =64, kernel_size=3, stride=1, padding=1
	MaxPool2d_2	kernel_size=2, stride=2
	AdaptiveAvgPool2d_2	output_size=1
	Linear_1	out_features=8
	Linear_2	out_features=1

tains a 3x3 convolutional layer with 64 filters, followed by a BN layer, ReLU layer, and maxpool layer. The final output is transformed to a range of 0 to 1 using two fully connected layers and a sigmoid function. The class with the highest score is the final classification result. In the SPA and CCA modules, we set $C' = C/8 = 256$.

Training and testing details. All experiments are implemented on Pytorch and a GeForce GTX 24GB 3090 GPU, Silver 4214R CPU. The optimizer is Adam [33]. The initial learning rate of 1.0e-5 and a decay learning rate of 1.0e-6. Besides a cosine scheduler for warm start. Training stops when the learning rate is less than 1.0e-7, with a patience of 7. We use the mean square error (MSE) loss when training, where positive sample labels are 0 and negative sample labels are 1. For each episode, we compute with 8 images from each category, i.e., K = 8. Therefore, the experiment is a 4-way 8-shot, and a total of 32 images (4 x 8) are used as a batch for training. In addition, to test the robustness, we perform different post-processing operations on the testing set. Details and experimental results are given in Section 4.3.

Performance metric. We use accuracy (ACC) to evaluate the experimental results, which is calculated using the formula $ACC = \frac{TP+TN}{P+N} \times 100\%$.. Here, TP and TN are the numbers of correctly classified positive and negative samples, respectively. P and N are the total number of positive and negative samples.

4.3. Comparisons with State-of-the-Art Works

In this section, several advanced methods have been selected for comparison. They are summarized below.

- Xception [12] performed well in GAN face detection experiments [55]. The optimizer is Adam [33]. The initial learning rate is set to 1.0e-5, and the learning rate decay is 1.0e-6. Training stops when the learning rate is less than 1.0e-7, with a patience of 7 and batch size is 16. (Xception)
- Yao et al. [65] used transfer learning and feature fusion module to generated image. The optimizer is SGD optimizer. The model was trained for 80 epochs with a batch size of 32. (CGNet)
- Ju et al. [29] combined global spatial information from the whole image and local informative features from multiple patches selected by an adaptation module. The optimizer is Adam [33]. The initial learning rate is set to 1.0e-5, and the learning rate decay is 1.0e-6. Training stops when the learning rate is less than 1.0e-7, with a patience of 7 and batch size is 32. (FGLNet)
- Li et al. [38] detected authenticity by Estimating Artifact Similarity. The optimizer is Adam [33]. The initial learning rate is set to 10e-5, then the model is trained for 200 epochs and optimized with Adam. (GaseNet)
- Chen et al. [7] proposed a method based on dual-color space to detect differences. The optimizer is Adam [33]. The initial learning rate is set to 1.0e-5, and the learning rate decay is 1.0e-6. Training stops when the learning rate is less than 1.0e-7, with a patience of 7 and batch size is 16. (Dual-Color)

We mix positive samples from different domains into one class and negative samples as well, i.e., CelebA-HQ and FFHQare mixed as one class of positive samples, and PGGAN and StyleGAN are mixed as one class of negative samples, and then binary detection is performed.

Results on original images. As shown in Table 3, in order to verify the in-sample datasets accuracy and generalization of the detection methods, we conduct comparative experiments on different GAN-generated face datasets (without any post-processing of the images). It can be seen that the proposed RENet outperforms the previous methods in both inner and out-of-sample datasets. RENet benefits from the cross-image relations of relational network, enabling better adaptive recognition of image relevance based on given sample examples. Additionally, it accurately determines the authenticity of images by contrasting the attentional differences between images. In Section 4.4, we will provide more detailed results of ablation and comparative experiments. Although GaseNet [38] is based on a relational network for detecting artifacts, the detection performance of RENet is better than GaseNet. The ROC curves of StyleGAN2 in Figure 5 also indicate that RENet can maintain excellent discriminative capability at different decision thresholds.

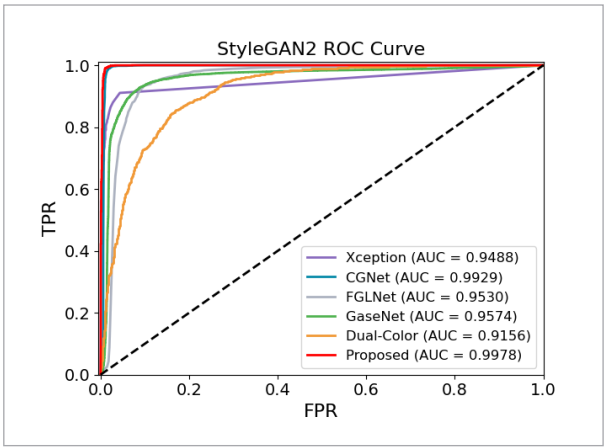
Furthermore, to better assess the performance of the models, we compared the inference times of each algorithm. As shown in Table 4, while FGLNet per-

Table 3

Detection accuracy of the different methods (%), the best results of the experiment in bold

Methods	in-sample datasets		out-of-sample datasets					
	ProGAN	StyleGAN	StyleGAN2	StarGAN	BeGAN	LsGAN	WgGANGP	RelGAN
Xception[57]	98.57	98.70	83.97	53.70	49.37	68.84	80.80	73.27
CGNet[67]	99.85	99.23	92.72	98.46	89.20	91.11	98.47	**99.94**
FGLNet[68]	99.01	98.85	85.63	**99.91**	**94.31**	96.52	99.03	95.68
GaseNet[19]	96.80	96.55	89.38	85.54	77.18	90.51	96.33	93.15
Dual-Color[13]	97.53	97.23	79.83	76.30	54.35	82.80	94.80	92.83
RENet (Proposed)	**99.93**	**99.43**	**94.83**	97.17	93.55	**98.20**	**99.35**	99.73

Figure 5
ROC curve for different methods in StyleGAN2



forms well in generalization, its inference time is increased due to the necessity of adaptively selecting local images for fusion, resulting in an excessive number of parameters. CGNet, on the other hand, sacrifices some generalization performance to improve overall efficiency. In contrast, RENet maintains excellent generalization performance while also preserving good inference times, making it more favorable for practical applications on real devices.

Results on robustness against post-processing operations. In reality, some criminals use GAN to generate faces maliciously and often post-process the images to evade detection algorithms. Therefore, we assessed the robustness of our system to several common attacks. They are JPEG compression (compression factors of 95, 90, 85, 80, 75, 70), Gaussian blur (kernel size and standard deviation of [3, 0.5], [3, 1.0], [3, 1.5], [5, 0.5], [5, 1.0], [5, 1.5]), resizing (scaling factors of 40, 60, 80, 120, 140, 160), Gaussian noise (standard deviation of 0.01, 0.02, 0.03, 0.04, 0.05), and Gamma correction (gamma values of 0.4, 0.6, 0.8, 1.2, 1.4, 1.6). Our method's performance was compared to other methods on both inner and out-of-sample datasets, as shown in Figure 6 and Figure 7, respectively. The results indicate that, in most cases, while the performance of other methods decreased significantly or

Table 4
Detection Average inference time (seconds) comparison of six methods for a picture with 256 × 256 resolution

Methods	Inference time
Xception [65]	0.026
CGNet [67]	0.029
FGLNet [68]	0.045
GaseNet [19]	0.030
Dual-Color [13]	0.052
RENet (Prposed)	0.032

Figure 6
The detection accuracy (%) of different methods on in-sample datasets for various post-processing

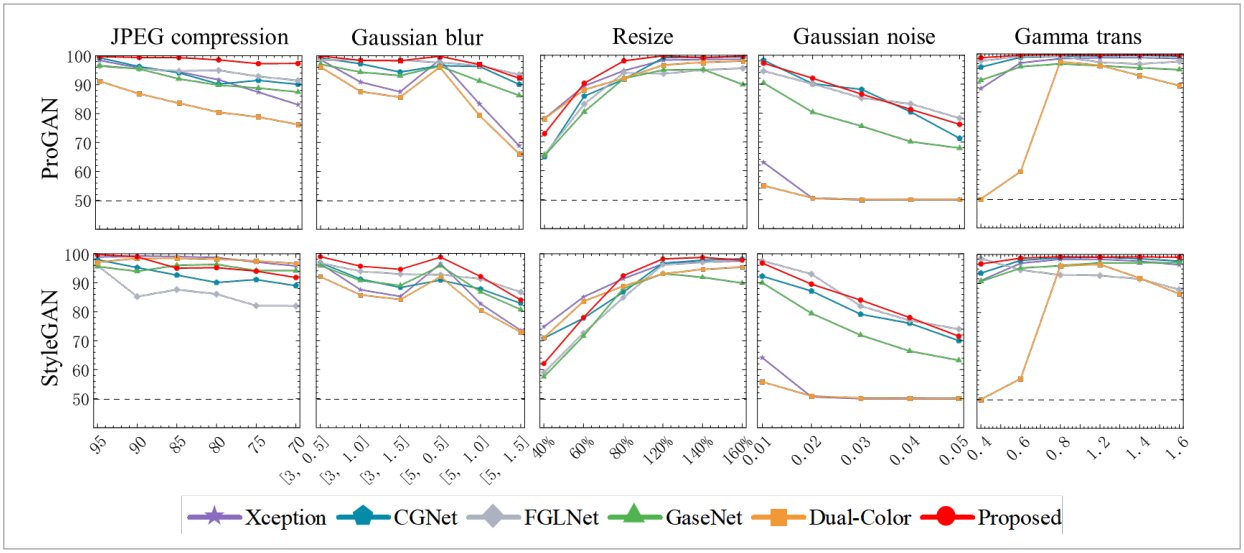
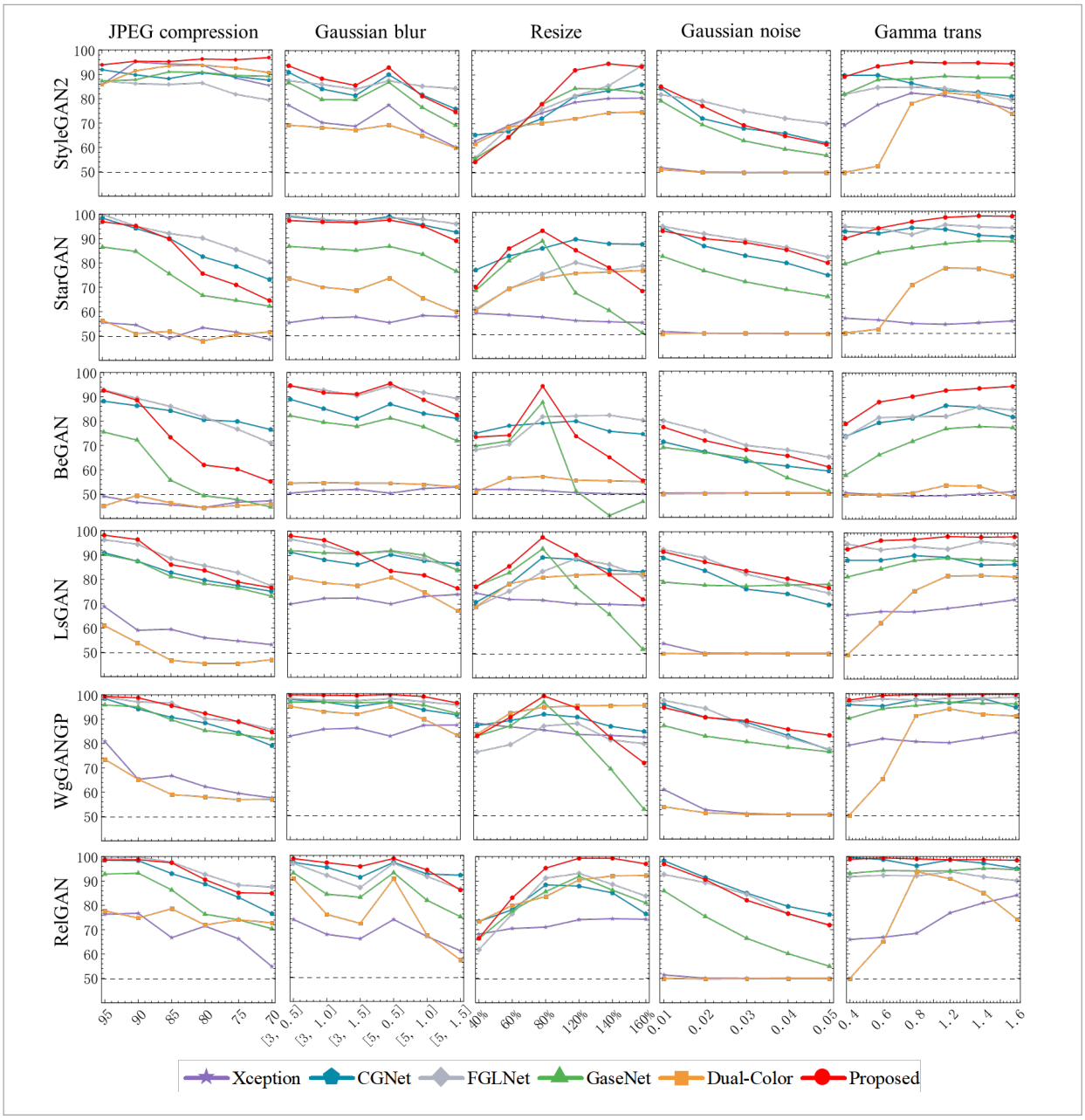


Figure 7

The detection accuracy (%) of different methods on out-of-sample datasets for various post-processing operations



became ineffective when attacked, our method maintained ideal detection performance.

This could be attributed to the fact that, even when post-processings cause degradation, the extracted features maintain a similarity to the prototype from the same category. Reliable detection results can be achieved through feature relation. Additionally, we visualized the attention points of RENet when processing post-processed images. As shown in Figure 8, it indicates that when the image is disturbed, RENet not only focuses on the intended areas but also pays attention to some background edges to assist in dis-

Figure 8
Activation maps from the proposed model after post-processing operations. From left to right, the column shows the images after the post-processed images through JPEG compression, Gaussian blur, resizing, Gaussian noise and Gamma trans

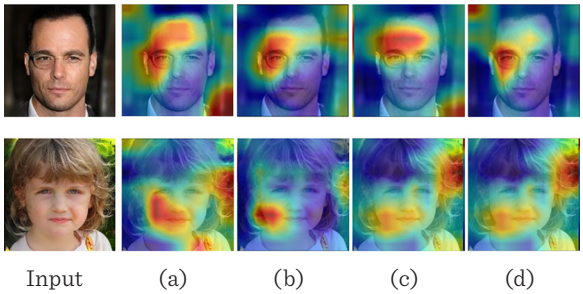
tinguishing authenticity, aligning with what has been suggested in [19]. Regarding the suboptimal performance of resizing on some datasets, we speculate that this is due to pixel loss caused by compressing all images uniformly to the same resolution before training.

4.4. Ablation Study and Selection of RENet Structure

In this section, we investigate four questions: (1) The impact of DSA and CCA modules on the effectiveness of RENet detection. (2) How does the RENet distinguish fake faces in out-of-sample datasets? (3) How to design the network architecture to optimize the performance of RENet. (4) What impact will batch size and learning rate have on the model?

For the first question, we have conducted ablation experiments in Table 5. The experimental results show that the DSA and CCA modules can improve the network's detection generalization ability, thereby verifying the effectiveness of the proposed modules. The visualization in Figure 9 demonstrates that the features extracted by the network become more precise and rational with the addition of the DSA module and CCA

Figure 9
The first column on the left is the input query set, and the red boxes represent the clearly fake areas. (a), (b), (c) and (d) are the feature attention maps without any module, with the DSA module, with the CCA module and with the DSA+CCA module, respectively



module separately. Especially when combining the two modules, the network's attention accuracy further improves. Besides, during the training process, as shown in Figure 10, the network converges more rapidly after integrating DSA and CCA. This indicates their successful role in enabling the network to perceive relevant semantic features at different positions, and facilitating a more accessible learning process for comparison.

Figure 10
The influence of DSA and CCA on the training loss

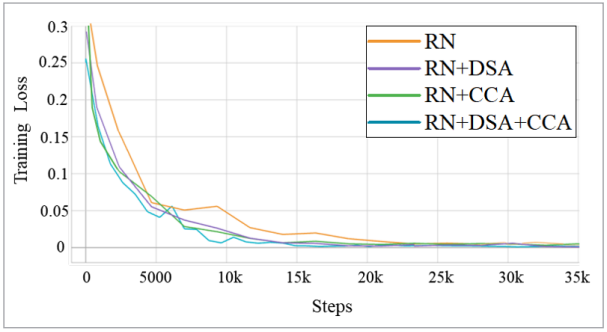


Table 5
Ablation experiments of the DSA and CCA modules (%)

	ProGAN	StyleGAN	StyleGAN2	StarGAN	BeGAN	LsGAN	WgGANGP	RelGAN
RN	98.6	96.1	93.1	94.2	90.6	97.3	97.6	97.9
RN + DSA	99.9	99.1	94.6	95.7	92.7	98.8	99.3	99.1
RN + CCA	99.5	98.6	93.2	95.4	91.5	93.7	98.6	98.4
RN + DSA + CCA	99.9	99.4	94.8	97.2	93.6	98.2	99.4	99.7

For the second question, Figure 11 visualizes the primary focus of RENet on out-of-sample datasets. It can be observed that, compared to detecting images within the dataset, when detecting images from out-of-sample datasets, the network tends to focus on detecting artifacts at the edges of faces and background hair. It may be due to differences in the strategies of different generation models, leading to variations in the correlation of facial texture features.

Figure 11

The visualization results of feature maps for out-of-sample datasets

For the third question, we propose the following verification directions: (1) After extracting features using the shared embedding network, should the multiple features of the support set be element-wise sum or element-wise avg, which is more suitable for GAN generated face detection? (2) How many kernels should the convolutional filtering have in the score network? (3) How do the convolutional layers in the score network affect detection ability?

To explore the above directions, we have made some modifications to RENet as shown in Figure 12. (1) RENet-1: Instead of element-wise sum, we use element-wise avg before feeding the support set features into the Feature Augmentation. (2) In RENet-2 and RENet-3, we modify the two convolutional kernels in the score network to 32 and 128, respectively. (3) In RENet-4, we add a convolutional layer to the score network, while in RENet-5, we reduce a convolutional layer in the score network. To make a fair comparison, we keep the other structures the same as the original RENet except for the corresponding modifications.

As shown in Table 6, We can draw the following conclusion. First, if we change element-wise sum to element-wise avg, the inner and out-of-sample datasets' detection accuracy of the network will decrease by 0.34% and 1.55%. This is attributed to the fact that element-wise sum effectively amplifies the distinctions and commonalities in the extracted features, facilitating the network in better contrasting the feature associations between the support set and query set. On the contrary, element-wise avg diminishes these differences, impacting the network's ability to discern their associations. Second, a convolutional kernel of 64 is more suitable for RENet, and a larger (128) or smaller (32) kernel only brings negative improvements, especially when the convolutional kernel is 128, the generalization accuracy decreases by 3.85%, indicating that an appropriately sized convolutional kernel can enhance the feature augmentation results of DSA. Third, using two convolutional layers before score mapping is more stable than using one or three

Figure 12

The complete RENet network is on the left, and various modified networks are within the dashed lines on the right. The red font indicates the modified parameters. (4) and (5) do not modify the parameters but modify the number of convolutional layers

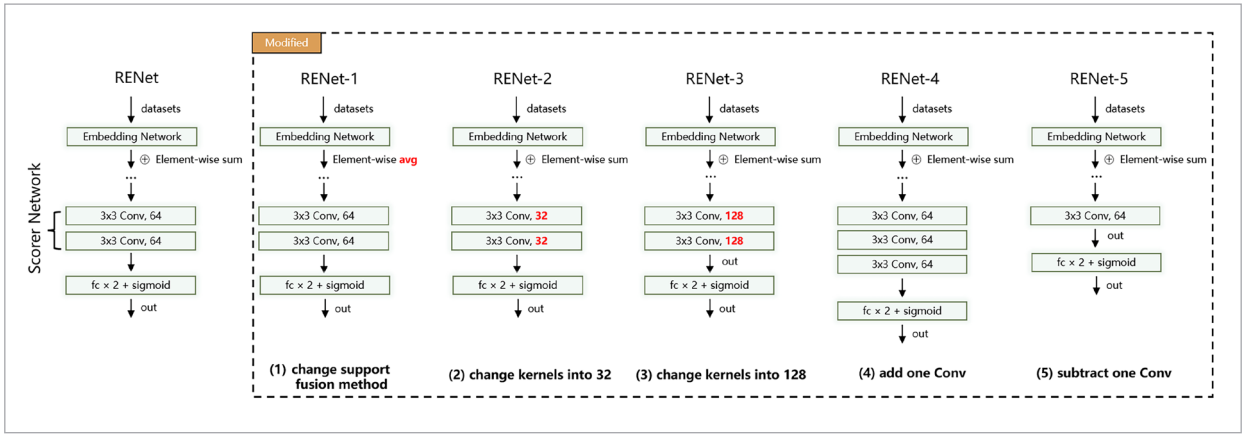


Table 6
The influence of RENet structure selection on accuracy (%)

Models		Description	in-sample datasets average accuracy	out-of-sample datasets average accuracy
Proposed	RENet	\	**99.68**	**96.85**
	RENet-1	change support fusion method into average	99.35 (-0.34)	95.30 (-1.55)
	RENet-2	change Score Network's kernels into 32	99.58 (-0.1)	95.29 (-1.56)
Modified	RENet-3	change Score Network's kernels into 128	98.99 (-0.69)	93.00 (-3.85)
	RENet-4	add one same conv in Score Network	99.06 (-0.62)	96.20 (-0.65)
	RENet-5	subtract one same conv in Score Network	99.11 (-0.57)	94.10 (-2.75)

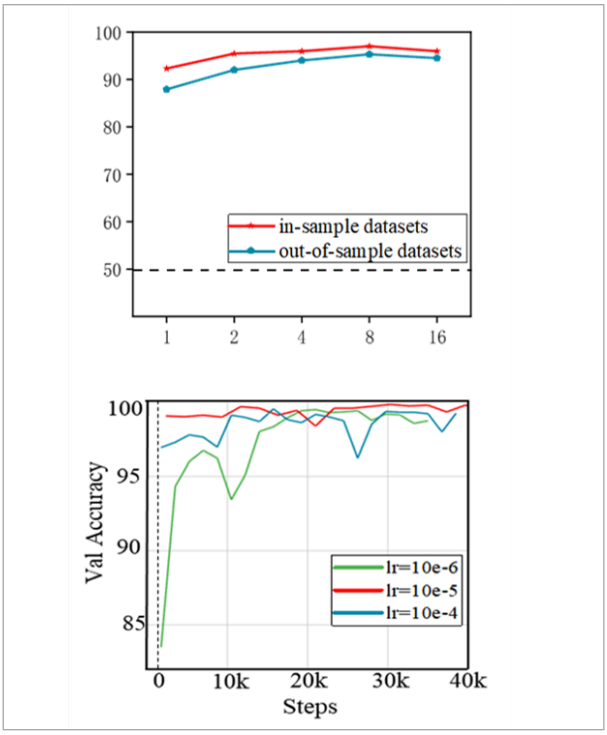
convolutional layers. Adding a convolutional layer not only reduces the detection accuracy, but also increases the network parameters, while reducing a layer cannot fully integrate the extracted features, among which the impact on generalization accuracy is the greatest.

For the fourth question, we conduct two experiments. The first one involved training the model with batch sizes of 1, 2, 4, 8 (in this paper), and 16, observing the accuracy on in-sample and out-of-sample datasets when the epoch equals 20. The second experiment set the initial learning rates to 10e-4, 10e-5 (in this paper), and 10e-6, monitoring the accuracy on the validation set at the same step. As shown in Figures 13 , appropriate batch sizes and learning rates can enhance detection accuracy, while excessively small or large batch sizes may impact generalization ability.

4.5. Generated Image Detection of Various Category

With the continuous evolution and application of datasets [5, 61], as well as the ongoing iterations of GAN models [3, 53, 55, 73, 74], the range of generated images has extended beyond faces. Training solely on CelebA-HQ and FFHQ may not comprehensively test the performance of the proposed method. To thoroughly analyze RENet's performance across different categories, we further expand the testing scope, evaluating our proposed method. We employ the experimental settings outlined by Wang et al. [60], which

Figure 13
Above are batch size ablation studies, and below are learning rate ablation studies on the RENet



crafted to show the generality of a generated image detector trained by a special GAN model in identifying other GAN models (not restricted to faces alone). Detailed information regarding the generated imag-

Figure 14
Some samples in the experimental datasets. The first row represents real images, and the second row corresponds to images generated by GAN

Table 7
Details of datasets. Generated images are not limited to faces alone

Family	Models	Source	Categories
Uncontiditonal GANs	ProGAN	LSUN	bottle, airplane, chair, horse, etc.
	StyleGAN	LSUN	car, cat, bedroom
	StyleGAN2	LSUN	car, cat, church, horse
	BigGAN	ImageNet	fish, snake, person, road, etc.
Contiditonal GANs	CycleGAN	Style/object transfer	apple, orange, zebra, winter, etc.
	StarGAN	CelebA	person
	GauGAN	COCO	bear, truck, spoon, sandwich, etc.
Deepfake	FaceForensics++	Videos of faces	face

es is presented in Figure 14 and Table 7. We choose Lsun [66] as the source for real images and select one category (horse) among the 20 classes generated by ProGAN as the fake images. Each category comprises 18,000 fake images and an equal number of real images. The training settings are the same as those mentioned in Section 4. In addition to Accuracy (ACC.), we also use the Average Precision (A.P.) as an evaluation metric. The A.P. is calculated using an alternative measurement method in method [60], which approximates the area under the precision-recall curve with the use of a few thresholds.

We first assessed the detection performance of each method across various categories within the in-sample datasets. As illustrated in Figure 15, even with training limited to a single category, our proposed RENet surpasses other methods in detecting unseen categories. Notably, in categories like Bus and Motorbike, where alternative methods show lower accuracy, RENet maintains an accuracy exceeding 90%. This suggests that proposed network is capable of effectively finding feature connections and semantic information within the same GAN model. Additionally, Figure 16 visualizes the attention of the proposed model on different types of images generated by the same GAN. We observed that, in this scenario, the model primarily focuses on discerning the authenticity by analyzing the content within the objects themselves.

Table 8 shows the comparison results between our proposed RENet and other methods. It can be observed that RENet demonstrates better generalization capabilities across most GAN models compared

Figure 15
Comparison of performance in unknown categories on images generated by ProGAN(%)

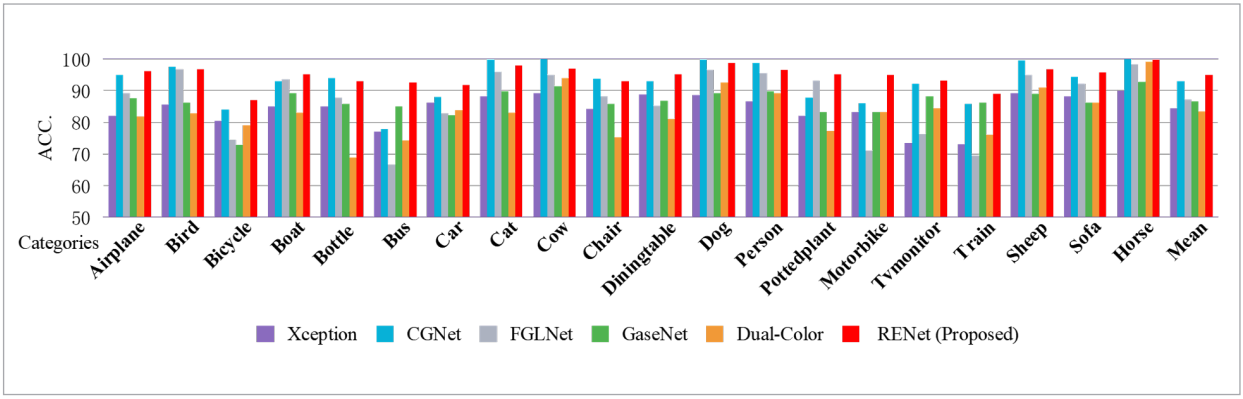


Figure 16
Activation maps in various categories after training on horse images generated by ProGAN

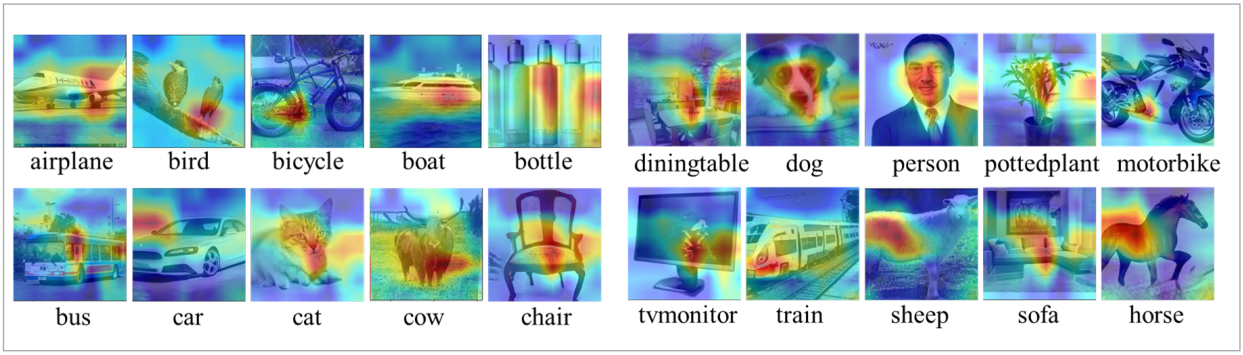


Table 8
Classification accuracy (%) with cross-model & unknow category

Methods	Test Models																	
	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.	ACC.	A.P.
Xception	84.3	89.5	74.3	80.1	77.3	87.6	58.1	57.8	60.4	67.1	98.7	99.9	58.0	59.0	54.9	59.0	70.8	75.0
CGNet	93.0	98.9	76.4	96.8	85.2	94.7	71.5	70.1	74.2	80.9	99.8	99.8	52.7	63.1	59.0	70.1	77.2	84.3
FGLNet	87.2	96.1	76.7	85.4	78.4	90.3	64.1	62.9	63.1	70.8	98.1	99.9	48.1	46.3	67.9	72.9	72.9	77.2
GaseNet	86.5	91.2	72.6	77.4	82.5	80.9	68.0	69.1	71.4	71.1	91.5	94.0	61.4	68.8	62.8	66.2	74.6	77.3
Dual-Color	83.3	95.1	70.1	83.8	77.5	85.5	58.2	60.6	63.5	72.3	84.3	95.6	56.3	55.1	57.8	62.1	68.9	76.3
RENet	94.9	96.2	78.4	81.8	81.6	84.4	71.6	72.2	76.0	80.4	97.8	98.7	63.2	60.5	69.3	69.5	78.1	80.4

to other algorithms. This is attributed to the enhanced relational network's ability to perceive semantic features and conduct comparisons. However, the generalization performance on GauGAN and Deepfake is not satisfactory. This happens due to GauGAN and ProGAN have similar semantic features, leading to overfitting problems in previous methods. In addition, the poor performance of the Deepfake model is due to the fact that it is not a GAN model and uses MSE loss and SSIM loss for training, resulting in a detection accuracy of only 69.3%.

5. Conclusion

In this work, we propose a relational embedding network called RENet for detecting GAN-generated face. It combines dual self-attention and cross-attention, enhancing both the relevant local features within an image and the global feature relationships between images. In addition, we observe that RENet can better generalize to unknown datasets by learning the structural correlations among features, and bring performance improvements to the network for detecting GAN-generated images. In further work, we will explore the effect of a relational network combined with an attention framework on different image forensics tasks.

Declarations

The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valued comments and suggestions. This work was supported by the projects of national social science foundation of China under Grant 21BXW077.

Data and Code Availability

The datasets that support the conclusions of this paper are cited in the article. These datasets were derived from the following public domain resources: https://github.com/NVlabs/ffhq-dataset; http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html; https://github.com/tkarras/progressive_growing_of_gans.

References

1. Barni, M., Kallas, K., Nowroozi, E., Tondi, B. CNN Detection of GAN-Generated Face Images based on Cross-Band Co-occurrences Analysis. In: 2020 IEEE International Workshop on Information Fo-rensics and Security (WIFS). Presented at the 2020 IEEE International Workshop on Infor-mation Forensics and Security (WIFS), 2020, 1-6. https://doi.org/10.1109/WIFS49906.2020.9360905
2. Berthelot, D., Schumm, T., Metz, L. BE-GAN: Boundary Equilibrium Generative Adversarial Networks, 2017. https://doi.org/10.48550/arXiv.1703.10717
3. Brock, A., Donahue, J., Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis, 2018. arXiv preprint arXiv:1809.11096.
4. Chai, L., Bau, D., Lim, S.-N., Isola, P. What Makes Fake Images Detectable? Understanding Proper-ties that Generalize, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision - ECCV 2020, Lecture Notes in Computer Science. Springer International Publishing, Cham, 103-120. https://doi.org/10.1007/978-3-030-58574-7_7
5. Chen, B., Ju, X., Xiao, B., Ding, W., Zheng, Y., de Albuquerque, V.H.C. Locally GAN-generated Face Detection Based on an Improved Xception. Information Sciences, 2021, 572, 16-28. https://doi.org/10.1016/j.ins.2021.05.006
6. Chen, B., Li, T., Wang, J., Zhao, G. GAN-Generated Face Detection with Strong Gener-alization Ability Based on Quaternions (in Chinese). Journal of Computer-Aided Design & Computer Graphics, 2022, 34, 734-742. https://doi.org/10.3724/SP.J.1089.2022.19015
7. Chen, B., Liu, X., Zheng, Y., Zhao, G., Shi, Y.-Q. A Robust GAN-Generated Face Detection Method Based on Dual-Color Spaces and an Improved Xception. IEEE Transactions on Circuits and Sys-tems for Video Technology, 2022, 32, 3527-3538. https://doi.org/10.1109/TCSVT.2021.3116679
8. Chen, B., Weijin, T., Wang, Y., Zhao, G. Distin-guishing Between Natural and GAN-Generated Face Images by Combining Global and Local Features. Chinese Journal of Electronics, 2022, 31, 59-67. https://doi.org/10.1049/cje.2020.00.372

9. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R. Local Relation Learning for Face Forgery Detec-tion. AAAI Conference on Artificial Intelligence, 2021. https://doi.org/10.1609/aaai.v35i2.16193
10. Cheng, H., Yang, S., Zhou, J. T., Guo, L., Wen, B. Frequency Guidance Matters in Few-Shot Learn-ing. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 11814-11824. https://doi.org/10.1109/ICCV51070.2023.01085
11. Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J. StarGAN: Unified Generative Adversari-al Networks for Multi-domain Image-to-Image Translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Pre-sented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, 2018, 8789-8797. https://doi.org/10.1109/CVPR.2018.00916
12. Chollet, F. Xception: Deep Learning with Depth-wise Separable Convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 1251-1258. https://doi.org/10.1109/CVPR.2017.195
13. Durall, R., Keuper, M., Keuper, J. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spec-tral Distributions. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Con-ference on Computer Vision and Pattern Recogni-tion (CVPR), IEEE, Seattle, WA, USA, 2020, 7887-7896. https://doi.org/10.1109/CVPR42600.2020.00791
14. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T. Leveraging Frequency Analysis for Deep Fake Image Recognition. In: Proceedings of the 37th International Conference on Machine Learning. Presented at the Interna-tional Conference on Machine Learning (ICML, PMLR), 2020, 3247-3258.
15. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H. Dual Attention Network for Scene Segmenta-tion. In: 2019 Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recogni-tion. Presented at the 2019 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2019. https://doi.org/10.1109/CVPR.2019.00326
16. Fu, T., Xia, M., Yang, G. Detecting GAN-generated Face Images Via Hybrid Texture and Sensor Noise Based Features. Multimedia Tools Applications, 2022, 81, 26345-26359. https://doi.org/10.1007/s11042-022-12661-1
17. Gao, S., Xia, M., Yang, G. Dual-Tree Complex Wavelet Transform-Based Direction Correlation for Face Forgery Detection. Security and Com-munication Networks, 2021, 1-10. https://doi.org/10.1155/2021/8661083
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets. In: Advances in Neural Information Processing Sys-tems. Presented at the Advances in Neural Infor-mation Processing Systems, Curran Associates, Inc., 2021.
19. Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., Verdoliva, L. Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-of-the-Art. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). Presented at the 2021 IEEE International Confer-ence on Multimedia and Expo (ICME), 2021, 1-6. https://doi.org/10.1109/ICME51207.2021.9428429
20. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C. Improved Training of Wasser-stein GANs. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017.
21. Guo, H., Hu, S., Wang, X., Chang, M.-C., Lyu, S. Robust Attentive Deep Neural Network for De-tecting GAN-Generated Faces. IEEE Access, 2022, 10, 32574-32583. https://doi.org/10.1109/ACCESS.2022.3157297
22. Guo, H., Hu, S., Wang, X., Chang, M.-C., Lyu, S. Eyes Tell All: Irregular Pupil Shapes Reveal GAN-Generated Faces. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Present-ed at the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Pro-cessing (ICASSP), IEEE, Singapore, Singapore, 2022, 2904-2908. https://doi.org/10.1109/ICASSP43922.2022.9746597
23. Guo, H., Hu, S., Wang, X., Chang, M.-C., Lyu, S. Eyes Tell All: Irregular Pupil Shapes Reveal GAN-Generated Faces. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Present-ed at the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Pro-cessing (ICASSP), IEEE, Singapore, Singapore, 2022, 2904-2908. https://doi.org/10.1109/ICASSP43922.2022.9746597
24. Guo, Z., Yang, G., Chen, J., Sun, X. Fake Face De-tection Via Adaptive Manipulation Traces Ex-traction Network. Computer Vision and Image Understanding, 2021, 204, 103170. https://doi.org/10.1016/j.cviu.2021.103170
25. Hassanin, M., Anwar, S., Radwan, I., Khan, F.S., Mian, A. Visual Attention Methods in Deep Learning: An In-Depth Survey, 2022.

26. He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

27. Hu, S., Li, Y., Lyu, S. Exposing GAN-Generated Faces Using Inconsistent Corneal Specular High-lights. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Pro-cessing (ICASSP). Presented at the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Toronto, ON, Canada, 2021, 2500-2504. https://doi.org/10.1109/ICASSP39728.2021.9414582

28. Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., Chu, W., Chen, J., Lin, W., Ma, K.-K. Cmua-watermark: A Cross-Model Universal Ad-versarial Watermark for Combating Deepfakes. In: Proceedings of the AAAI Conference on Arti-ficial Intelligence, 2022, 989-997. https://doi.org/10.1609/aaai.v36i1.19982

29. Jia, G., Zheng, M., Hu, C., Ma, X., Xu, Y., Liu, L., Deng, Y., He, R. Inconsistency-Aware Wavelet Dual-Branch Network for Face Forgery Detection. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3, 308-319. https://doi.org/10.1109/TBIOM.2021.3086109

30. Ju, Y., Jia, S., Ke, L., Xue, H., Nagano, K., Lyu, S. Fusing Global and Local Features for Gener-alized AI-Synthesized Image Detection. In: 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022, 3465-3469. https://doi.org/10.1109/ICIP46576.2022.9897820

31. Karras, T., Aila, T., Laine, S., Lehtinen, J. Progres-sive Growing of GANs for Improved Quality, Sta-bility, and Variation, 2018. https://doi.org/10.48550/arXiv.1710.10196

32. Karras, T., Laine, S., Aila, T. A Style-Based Gen-erator Architecture for Generative Adversarial Networks. In: Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recogni-tion. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 4401-4410. https://doi.org/10.1109/CVPR.2019.00453

33. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2020, 8107-8116. https://doi.org/10.1109/CVPR42600.2020.00813

34. Kingma, D.P., Ba, J. Adam: A Method for Sto-chastic Optimization, 2017. arXiv Preprint arXiv:1412.6980.

35. Laurinavičius, D., Maskeliūnas, R., Damaševičius, R. Improvement of Facial Beauty Prediction Using Artificial Human Faces Generated by Generative Adversarial Network. Cognitive Computing, 2023, 15, 998-1015. https://doi.org/10.1007/s12559-023-10117-8

36. Le, B.M., Woo, S.S. ADD: Frequency Attention and Multi-View Based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Imag-es. AAAI Conference on Artificial Intelligence, 2021.

37. Li, H., Li, W., Wang, S. Discovering and Incorpo-rating Latent Target-Domains for Domain Adap-tation. Pattern Recognition, 2020, 108, 107536. https://doi.org/10.1016/j.patcog.2020.107536

38. Li, J., Xie, H., Jiahong Li, Wang, Z., Zhang, Y. Frequency-Aware Discriminative Feature Learn-ing Supervised by Single-Center Loss for Face Forgery Detection. In: 2021 IEEE/CVF Confer-ence on Computer Vision and Pattern Recognition (CVPR). Presented at the 2021 IEEE/CVF Con-ference on Computer Vision and Pattern Recogni-tion (CVPR), IEEE, Nashville, TN, USA, 2021, 6454-6463. https://doi.org/10.1109/CVPR46437.2021.00639

39. Li, W., He, P., Li, H., Wang, H., Zhang, R. Detec-tion of GAN-Generated Images by Estimating Ar-tifact Similarity. IEEE Signal Processing Letters, 2022, 29, 862-866. https://doi.org/10.1109/LSP.2021.3130525

40. Li, Wei, Zhong, X., Shao, H., Cai, B., Yang, X. Multi-Mode Data Augmentation and Fault Diag-nosis of Rotating Machinery Using Modified AC-GAN Designed with New Framework. Ad-vanced Engineering Informatics, 2022, 52, 101552. https://doi.org/10.1016/j.aei.2022.101552

41. Liang, B., Wang, Z., Huang, B., Zou, Q., Wang, Q., Liang, J. Depth Map Guided Triplet Network for Deepfake Face Detection. Neural Networks, 2023, 159, 34-42. https://doi.org/10.1016/j.neunet.2022.11.031

42. Lin, H., Luo, W., Wei, K., Liu, M. Improved Xcep-tion with Dual Attention Mechanism and Feature Fusion for Face Forgery Detection. In: 2022 4th International Conference on Data Intelligence and Security (ICDIS). Presented at the 2022 4th Inter-national Conference on Data Intelligence and Se-curity (ICDIS), 2022, 208-212. https://doi.org/10.1109/ICDIS55630.2022.00039

43. Lin, Y.-J., Wu, P.-W., Chang, C.-H., Chang, E., Liao, S.-W. RelGAN: Multi-Domain Im-age-to-Image Trans-

lation via Relative Attributes. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, 5913-5921. https://doi.org/10.1109/ICCV.2019.00601

44. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Hui, X., Weiming, Z., Nenghai, Y. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 772-781. https://doi.org/10.1109/CVPR46437.2021.00083
45. Liu, Z., Luo, P., Wang, X., Tang, X. Deep Learning Face Attributes in the Wild, in: Proceedings of the IEEE International Conference on Computer Vi-sion. Presented at the Proceedings of the IEEE In-ternational Conference on Computer Vision, 2015, 3730-3738. https://doi.org/10.1109/ICCV.2015.425
46. Liu, Z., Qi, X., Torr, P. H. S. Global Texture En-hancement for Fake Face Detection in the Wild. In: 2020 IEEE/CVF Conference on Computer Vi-sion and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2020, 8057-8066. https://doi.org/10.1109/CVPR42600.2020.00808
47. Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., Paul Smolley, S. Least Squares Generative Adversarial Networks. In: Proceedings of the IEEE Interna-tional Conference on Computer Vision, 2017, 2794-2802. https://doi.org/10.1109/ICCV.2017.304
48. Matern, F., Riess, C., Stamminger, M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In: 2019 IEEE Winter Applica-tions of Computer Vision Workshops (WACVW). Presented at the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE, Waikoloa Village, HI, USA, 2019, 83-92. https://doi.org/10.1109/WACVW.2019.00020
49. Miao, C., Tan, Z., Chu, Q., Liu, H., Hu, H., Yu, N. F 2 Trans: High-Frequency Fine-Grained Trans-former for Face Forgery Detection. IEEE Transac-tions on Information Forensics and Security, 2023, 18, 1039-1051. https://doi.org/10.1109/TIFS.2022.3233774
50. Nataraj, L., Mohammed, T. M., Manjunath, B. S., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K. Detecting GAN Generated Fake Images Using Co-occurrence Matrices. ei 31, 2019, 532-1-532-7. https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-532
51. Neekhara, P., Hussain, S., Zhang, X., Huang, K., McAuley, J., Koushanfar, F. FaceSigns: Semi-Fragile Neural Watermarks for Media Au-thentication and Countering Deepfakes, 2022. arXiv Preprint arXiv:2204.01960.
52. Ouyang, J., Huang, J., Wen, X., Shao, Z. A Semi-Fragile Watermarking Tamper Localization Method Based on QDFT and Multi-View Fusion. Multimedia Tools Application, 2022. https://doi.org/10.1007/s11042-022-13938-1
53. Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., Huang, G. On the Integration of Self-Attention and Convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 815-825. https://doi.org/10.1109/CVPR52688.2022.00089
54. Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y. Se-mantic Image Synthesis with Spatially-Adaptive Normalization. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 2337-2346. https://doi.org/10.1109/CVPR.2019.00244
55. Rhee, H., Jang, Y. I., Kim, S., Cho, N. I. LC-FDNet: Learned Lossless Image Compression with Fre-quency Decomposition Network. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Or-leans, LA, USA, 2022, 6023-6032. https://doi.org/10.1109/CVPR52688.2022.00594
56. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M. Face Forensics: Learn-ing to Detect Manipulated Facial Images, in: Pro-ceedings of the IEEE/CVF International Confer-ence on Computer Vision. Presented at the Pro-ceedings of the IEEE/CVF International Confer-ence on Computer Vision, 2019, 1-11. https://doi.org/10.1109/ICCV.2019.00009
57. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., Hospedales, T. M. Learning to Compare: Rela-tion Network for Few-Shot Learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, 2018, 1199-1208. https://doi.org/10.1109/CVPR.2018.00131
58. Vasudeva B., Deora P., Bhattacharya S., Pradhan P.M. Compressed Sensing MRI Reconstruction With Co-VeGAN: Complex-Valued Generative Adversarial Network. Presented at the Proceed-ings of the IEEE/CVF Winter Conference on Ap-plications of Computer Vision, 2022, 672-681. https://doi.org/10.1109/WACV51458.2022.00184

59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. Attention is All you Need, in: Advances in Neural Information Processing Systems. Curran Associ-ates, Inc, 2017.

60. Wang, J., Zeng, K., Ma, B., Luo, X., Yin, Q., Liu, G., Jha, S. K. GAN-generated Fake Face Detection via Two-stream CNN with PRNU in the Wild. Multimedia Tools Application, 2022, 1-19. https://doi.org/10.1007/s11042-021-11592-7

61. Wang, S.-Y., Wang, O., Zhang, R., Owens, A., Efros, A. A.. CNN-Generated Images Are Surpris-ingly Easy to Spot for Now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 8695-8704. https://doi.org/10.1109/CVPR42600.2020.00872

62. Wei, W., Ho, E.S.L., McCay, K. D., Damaševičius, R., Maskeliūnas, R., Esposito, A. Assessing Facial Symmetry and Attractiveness Using Augmented Reality. Pattern Analysis and Applications, 2022, 25, 635-651. https://doi.org/10.1007/s10044-021-00975-z

63. Wu, H., Kuo, H.-C., Zheng, N., Hung, K.-H., Lee, H.-Y., Tsao, Y., Wang, H.-M., Meng, H. Partially Fake Audio Detection by Self-Attention-Based Fake Span Discovery. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, 9236-9240. https://doi.org/10.1109/ICASSP43922.2022.9746162

64. Wu, Z., Li, Y., Guo, L., Jia, K. Parn: Posi-tion-aware relation Networks for Few-shot Learning. In: Proceedings of the IEEE/CVF Inter-national Conference on Computer Vision, 2019, 6659-6667. https://doi.org/10.1109/ICCV.2019.00676

65. Yang, X., Li, Y., Lyu, S. Exposing Deep Fakes Using Inconsistent Head Poses. In: ICASSP 2019 - 2019 IEEE International Conference on Acous-tics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2019-2019 IEEE Interna-tional Conference on Acoustics, Speech and Sig-nal Processing (ICASSP), IEEE, Brighton, United Kingdom, 2019, 8261-8265. https://doi.org/10.1109/ICASSP.2019.8683164

66. Yao, Y., Zhang, Z., Ni, X., Shen, Z., Chen, L., Xu, D. CGNet: Detecting Computer-generated Images Based on Transfer Learning with Attention Mod-ule. Signal Processing: Image Communication, 2022, 105, 116692. https://doi.org/10.1016/j.image.2022.116692

67. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2015. arXiv preprint arXiv:1506.03365.

68. Yuchen, L., Yon, Z., Junchi, Y., Wei, L. General-izing Face Forgery Detection with High-Frequency Features. Presented at the Pro-ceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition, 2021, 16317-16326.

69. Yuyang, Q., Guojun, Y., Lu, S., Zixuan, C., Jing, S. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues, in: Andrea Vedaldi, Horst Bischof, Thomas Brox, Jan-Michael Frahm (Eds.), European Conference on Computer Vision-ECCV 2020, Lecture Notes in Computer Science. Springer International Pub-lishing, Cham, 2020, 86-103. https://doi.org/10.1007/978-3-030-58610-2_6

70. Zheng, Q., Tian, X., Yu, Z., Jiang, N., Elhanashi, A., Saponara, S., Yu, R. Application of Wave-let-packet Transform Driven Deep Learning Method in PM2.5 Concentration Prediction: A Case Study of Qingdao, China. Sustainable Cities and Society, 2023, 92, 104486. https://doi.org/10.1016/j.scs.2023.104486

71. Zheng, Q., Zhao, P., Li, Y., Wang, H., Yang, Y. Spectrum Interference-Based Two-Level Data Augmentation Method in Deep Learning for Au-tomatic Modulation Classification. Neural Com-puting & Applications, 2021, 33, 7723-7745. https://doi.org/10.1007/s00521-020-05514-1

72. Zheng, Q., Zhao, P., Wang, H., Elhanashi, A., Sa-ponara, S. Fine-Grained Modulation Classifica-tion Using Multi-Scale Radio Transformer with Dual-Channel Representation. IEEE Communica-tion Letters, 2022, 26, 1298-1302. https://doi.org/10.1109/LCOMM.2022.3145647

73. Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S. Detecting Camouflaged Object in Frequency Domain In: 2022 IEEE/CVF Conference on Com-puter Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, 2022, 4494-4503. https://doi.org/10.1109/CVPR52688.2022.00446

74. Zhu, J.-Y., Park, T., Isola, P., Efros, A. A. Unpaired Image-to-Image Translation Using Cy-cle-Consistent Adversarial Networks. In: 2017 IEEE International Conference on Computer Vi-sion (ICCV). Presented at the 2017 IEEE Interna-tional Conference on Computer Vision (ICCV), IEEE, Venice, 2017, 2242-2251. https://doi.org/10.1109/ICCV.2017.244

75. Zou, H., Ak, K.E., Kassim, A. A. Edge-Gan: Edge Conditioned Multi-View Face Image Generation. In: 2020 IEEE International Conference on Image Processing (ICIP). Presented at the 2020 IEEE In-ternational Conference on Image Processing (ICIP), 2020, 2401-2405. https://doi.org/10.1109/ICIP40778.2020.9190723