Mining Generalized Features for Detecting AI-Manipulated Fake Faces
Yang Yu, Rongrong Ni and Yao Zhao, Senior Member, IEEE

Abstract—Recently, AI-manipulated face techniques have developed rapidly and constantly, which has raised new security issues in society. Although existing detection methods consider different categories of fake faces, the performance on detecting the fake faces with "unseen" manipulation techniques is still poor due to the distribution bias among cross-manipulation techniques. To solve this problem, we propose a novel framework that focuses on mining intrinsic features and further eliminating the distribution bias to improve the generalization ability. Firstly, we focus on mining the intrinsic clues in the channel difference image (CDI) and spectrum image (SI) from the camera imaging process and the indispensable step in AI manipulation process. Then, we introduce the Octave Convolution (OctConv) and an attention-based fusion module to effectively and adaptively mine intrinsic features from CDI and SI. Finally, we design an alignment module to eliminate the bias of manipulation techniques to obtain a more generalized detection framework. We evaluate the proposed framework on four categories of fake faces datasets with the most popular and state-of-the-art manipulation techniques, and achieve very competitive performances. To further verify the generalization ability of the proposed framework, we conduct experiments on cross-manipulation techniques, and the results show the advantages of our method.

Index Terms—AI-manipulated face detection, intrinsic features mining, attention fusion, generalization ability.

I. INTRODUCTION

HUMAN face images and videos contain personal information and play an important role in daily life, such as communication, access control and payment. However, with the remarkable development of AI-manipulated techniques, it is becoming increasingly easy to produce fake faces. Unlike previous simple face manipulation techniques (e.g., splicing), AI-manipulated techniques can easily produce more realistic fake face images, even fake face videos. Specifically, these techniques can synthesize non-existent face images or directly manipulate face expressions images, face attributes images, even manipulate identities in videos. Fig. 1 presents four categories of fake faces with various AI-manipulated techniques, by which people can be easily fooled. These realistic fake faces may be abused for malicious purpose, raising security and privacy issues in our society. More critically, the detection methods of previous simple face manipulation techniques cannot work on the AI-manipulated fake faces. Therefore, it is extremely necessary to develop effective methods for detecting AI-manipulated fake faces.

The security concerns have motivated a number of studies for detecting AI-manipulated fake faces. Some hand-crafted features-based detection methods heavily depend on the specific defects in the manipulation process and cannot extract intrinsic features (e.g., eye color variance and lack of eye blinking), which determines the short-term effectiveness of these methods. Some simple learning-based detection methods aim to deal with certain categories of manipulation faces, but their effectiveness is limited to the certain categories they are trained for. To further consider the detection of all categories of fake faces, some learning-based methods focus on finding relatively common variances (e.g., neuron behaviors variance) between real and fake faces and taking advantage of incremental learning or domain adaptation to detect fake faces continuously. Nevertheless, due to the distribution bias among cross-manipulation techniques, the above detection methods perform poorly when applied to the fake faces produced by "unseen" manipulation techniques, as shown in Fig. 2. Thus, how to improve the generalization ability on detecting the fake faces with unseen manipulation techniques becomes increasingly challenging. To address this issue, several attempts have been proposed recently. Some approaches attempt to simulate and generalize to unseen fake face generators with one specific architecture (e.g., AutoGAN and ProGAN). Some approaches attempt to utilize the attention mechanism to highlight the informative regions to reduce the interference of bias mentioned above. Some approaches assume the existence of a blending step in the fake face manipulation process (e.g., Face X-ray). However, since the intrinsic features cannot be captured...
well enough and the assumption is too strong, the above methods cannot perform well on the fake faces produced by new emerging manipulation techniques.

In this paper, we aim to improve the generalization ability on detecting fake faces with unseen manipulation techniques. As analyzed above, there are still two important issues on improving generalization ability. On one hand, the intrinsic features in all categories of fake faces produced by various manipulation techniques still need to be mined. On the other hand, how to eliminate the bias of manipulation techniques becomes increasingly pivotal. In view of this, we design a novel framework that focuses on mining the intrinsic features and further reducing the bias mentioned above to obtain a more generalized framework. Firstly, for fake faces with various AI-manipulated techniques, we mine two intrinsic clues in the channel difference image (CDI) and spectrum image (SI) view of the camera imaging process and the indispensable step in AI manipulation process, rather than depending on the specific defects in the manipulation process. In particular, we find both clues are related to frequency domain information, thus the Octave Convolution (OctConv) which has been proved to be efficient for capturing frequency information is employed to learn intrinsic features for detecting AI-manipulated fake faces from CDI and SI. Moreover, an attention-based fusion module is exploited to adaptively weight features, guiding the effective performance of fused features. Finally, to obtain a more generalized framework, we design an alignment module to eliminate the distribution bias among cross-manipulation techniques.

Our main contributions can be summarized as:

- We mine intrinsic clues from the camera imaging process and the indispensable step in AI manipulation process and further adopt OctConv and an attention-based fusion module to effectively mine intrinsic features for detecting AI-manipulated fake faces.
- To further improve generalization ability for our framework, we design an alignment module to reduce the bias by minimizing the difference in feature distribution among cross-manipulation techniques.
- We conduct extensive evaluations on four categories of fake faces datasets as well as our proposed cross-manipulation technique based protocols to verify the comprehensiveness and generalization ability of our detection framework. The results show the effectiveness of our proposed framework compared with other state-of-the-art methods, especially on the generalization performance. Extensive ablation studies demonstrate the effectiveness of each component in our framework.

The rest of the paper is organized as follows. Related works are briefly reviewed in Section II. Analysis of intrinsic clues and the proposed framework for mining generalized features to detect AI-manipulated fake faces are presented in Section III. Section IV shows the experimental results and corresponding analysis. The conclusion and future studies are drawn in Section V.

II. RELATED WORKS

In this section, we discuss the most relevant methods including fake face manipulation techniques and fake face detection methods.

A. Fake Face Manipulation Techniques

Existing AI-manipulated fake faces can be roughly categorized into four categories: entire face synthesis, facial expression manipulation, facial attribute manipulation, and face identity manipulation. For entire face synthesis, various Generative Adversarial Networks (GANs) are usually used to create entire non-existent faces. Karras et al. proposed StyleGAN [1] as an improved version of their previous popular approach ProGAN [2], which introduced an alternative generator architecture to synthesize highly varied and higher-quality human faces. Then they presented the StyleGAN2 [3] to further improve the quality of the generated face images. For facial expression manipulation, powerful GANs and 3D face reconstruction
methods are widely used for this manipulation. Ding et al. [4] proposed ExprGAN for photo-realistic facial expression manipulation with controllable expression intensity. Albert et al. [5] introduced a GAN conditioning scheme GANimation based on Action Units (AU) annotations, which manipulated human expression in a continuous manifold. Chen et al. [6] proposed HomInterpGAN to generate high-quality results for unpaired facial expression translation. Thies et al. [7] presented Face2Face for real-time facial expressions reenactment from one person to another via re-render and animation methods. For facial attribute manipulation, this manipulation edits single or multiple attributes in a face (e.g., gender, age, skin color, hair, and glasses), usually through GAN-based frameworks for general image translations and manipulations. Choi et al. [8] proposed StarGAN to perform image-to-image translations for multiple domains using a conditional attribute transfer network and achieved good visual results. He et al. [9] proposed AttGAN which provided realistic attribute manipulation results with other facial details well preserved by applying the attribute classification constraints. Liu et al. [10] proposed STGAN to improve the attribute manipulation ability and the image quality by incorporating selective transfer units with encoder-decoder. FaceApp1 popularized facial attribute manipulation as a consumer-level application, which provided 28 filters to modify specific attributes. For face identity manipulation, this manipulation replaces the face of one person with the face of another person. Two different approaches are usually considered: classical computer graphics-based techniques such as FaceSwap [11], and novel deep learning techniques known as DeepFakes [12] and NeuralTextures [13].

B. Fake Face Detection Methods

Fake face manipulation detection methods can be broadly classified into two classes, i.e., handcrafted features-based methods and deep learning-based methods.

The handcrafted features-based methods try to highlight specific defects in the fake face manipulation process. Matern et al. [13] detected DeepFakes and Face2Face videos based on visual artifacts, such as eye color variance, unconvincing specular reflections and missing details in the eye and teeth areas. Li et al. [19] proposed to expose AI-created fake videos by detecting the lack of eye blinking of the DeepFake videos. Then, they also utilized face warping artifacts, face landmark locations, and inconsistent head pose to expose DeepFake videos. Fernandes et al. [38] revealed DeepFake based on the lack of variations induced by heart beating. McCluskey et al. [39] analyzed that the GAN-generated images lack saturated regions. These detection methods rely on the handcrafted features which mostly depend on the specific defects in the manipulation process, thus the main drawback of these approaches remains that they are likely to soon become ineffective as generation methods evolve. Therefore, the intrinsic features of AI-manipulated fake faces still need to be mined.

For the learning-based methods, some simple CNN-based methods were proposed primarily. Afchar et al. [20] detected DeepFakes and Face2Face videos via two networks (Meso-4 and MesoInception-4) with a low number of layers that focus on the mesoscopic properties of images. Yu et al. [21] presented a method to detect fake images by learning GAN artificial fingerprints. Rossler et al. [22] introduced a face manipulation dataset FaceForensics++ and utilized Xception [40] to improve forgery detection accuracy in the presence of strong compression. Nguyen et al. [23] introduced a capsule network to detect forged images and videos. The recurrent neural network [24, 25] and the optical flow [26] were adopted to utilize the time information for exposing fake face videos. Mi et al. [41] equipped the algorithm with a much better comprehension of the global information with the self-attention mechanism to detect entire fake images. Liu et al. [42] presented Gram-Net that leveraged global image texture representations for generalization ability promotion on detecting the fake faces. He et al. [43] employed the ensemble of deep representations from multi color spaces for detecting fake images and further applied the random forest classifier against different post-processing attacks. To further consider the detection ability on all categories of fake faces, Wang et al. [27] introduced FakeSpotter to spot AI-manipulated fake faces by monitoring neuron behaviors. In order to continuously detect fake faces, Marra et al. [28] proposed an incremental learning detection method and Cozzolino et al. [29] introduced an autoencoder-based ForensicTransfer to detect fake face with novel manipulation techniques using a few examples, without worsening the performance on the previous ones. Qian et al. [44] proposed F$^3$-Net to detect fake faces based on two different but complementary frequency-aware clues. However, the effectiveness of these methods was limited to the manipulation techniques they were trained for, and most of these detectors perform poorly on the unseen manipulation techniques.

Some approaches focus on detecting the fake faces with unseen manipulation techniques. Xuan et al. [45] proposed a pre-processing step to reduce low level artifacts of GAN images and force the discriminator to learn more general forensic features for improving the generalization ability. Li et al. [33] found that most existing face manipulation techniques shared a common blending operation, thus proposed the Face x-ray to focus on the boundary of forged faces instead of the type of manipulation. Zhang et al. [30] proposed a GAN simulator to reproduce common GAN-image artifacts, which manifested as spectral peaks in the Fourier domain. Li et al. [46] observed that the fake images were more distinguishable from real ones in the chrominance components, especially in the residual domain. Dang et al. [32] utilized an attention mechanism to process and improve the feature maps for the fake face manipulation detection task. Wang et al. [31] detected GAN-generated images using a universal detector with careful pre and post-processing and data augmentation. Chai et al. [47] used a patch-based classifier with limited receptive fields to visualize which regions of fake images were more easily detectable and further showed a technique to exaggerate these detectable properties. However, these methods cannot work well on the fake faces with the state-of-the-art manipulation methods.

1https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341
III. PROPOSED METHOD

In this section, we illustrate the proposed framework for mining generalized features to detect AI-manipulated fake faces in detail. The pipeline of this framework is shown in Fig. 3. Firstly, we mine the two intrinsic clues from the CDI and SI. For more details and the theoretical analysis, refer to the following Section III-A. In Section III-B, we introduce the features learning module that perceives these two sources of information to mine intrinsic features. Subsequently, the attention-based feature fusion module that adaptively fuse features is described. In Section III-C, to further obtain a generalized framework, we specially propose an alignment module to eliminate the bias among cross-manipulation techniques in feature distribution.

A. Intrinsic Clues Analysis

1) CDI Analysis: In this part, we focus on mining the clues from the camera imaging process. In natural images, the high-frequency components across different color channels are highly mutually correlated and approximately equal due to the CFA interpolation algorithm during imaging process [48].

Hence, we conduct the following analysis: one color channel can be formally described as:

\[ I_c = I_c^l + I_c^h \]

where \( c \in \{R, G, B\} \), and \( h \) and \( l \) denote the high-frequency and low-frequency components of image color channels.

The channel difference can be expressed as:

\[ I_{c1} - I_{c2} = I_{c1}^l + I_{c1}^h - I_{c2}^l - I_{c2}^h \]

where \( c_1 \) and \( c_2 \) represent different color channels.

For real faces, due to the similarity of high-frequency components, \( I_{c1}^h \approx I_{c2}^h \). Therefore, the channel difference can be denoted as:

\[ I_{c1} - I_{c2} \approx I_{c1}^l - I_{c2}^l \]

As we can observe from Eq. (3), in the channel difference of real faces, the corresponding high-frequency components are filtered out, and only the low-frequency components are retained.

For entire-synthesis fake faces, there is no CFA interpolation algorithm in the generation process, which is different from...
the camera imaging process, thus the correlation of high-frequency components in color channels does not exist. For other categories of fake faces, because the change of values in three channels are different after the manipulation operation, thus the correlation of high-frequency components in color channels is destroyed. In both cases, the channel difference can not be expressed as Eq. (3). Therefore, the channel difference of four categories of fake faces contains more high-frequency components compared to real faces. Four examples are shown in Fig. 4 to verify this analysis. We calculate the channel difference images (CDI) of $R - G$ of real faces and four categories of fake faces, respectively. In all types of fake faces, the CDIs of $R - G$ contain more face details than the real ones. Therefore, the CDI contains discriminative intrinsic information for detecting fake faces.

2) SI Analysis: In this part, we focus on mining clues from the indispensable step in the process of generating AI fake faces. We study the pipeline of producing four categories of fake faces and find out the up-sampling modules are consistent. Zhang et al. [30] show that the up-sampling results in replications of spectra in the frequency domain, thus we mine intrinsic clues from the frequency spectrum. We check the average frequency spectrum obtained by Discrete Fourier Transform (DFT) from four types of fake faces and corresponding real faces to study the artifacts. For each manipulation technique, we choose 2000 face images/frames randomly. Compared to the real faces, we find that there are obviously grid-like patterns in the frequency spectrum of fake faces generated by different manipulation techniques, as shown in Fig. 5. The reason is that up-sampling actually replicate multiple copies in the spectrum of low-resolution images over high-frequency parts in the spectrum of final high-resolution images. In addition, the DeepFake faces lose high-frequency information compared to genuine ones due to interference of more pre- and post-processing in the video forgery process, hence the spectrum contains less high-frequency components, as shown in Fig. 5 (d). Therefore, the spectrum image (SI) also contains discriminative intrinsic clues for detecting fake faces.

According to above analysis, the discriminative clues in the CDI and SI rely on the fundamental differences between the camera imaging process and the AI-manipulated process, instead of depending on the specific defects in the manipulation process. Consequently, the CDI and SI could be used for mining intrinsic features for improving the generalization ability on detecting AI-manipulated fake faces.

Fig. 4. The real face images and the corresponding $R - G$ CDIs (left), and four categories of fake faces with different manipulation techniques and the corresponding $R - G$ CDIs (right). (a) Entire Synthesis. (b) Expression Manipulation. (c) Attribute Manipulation. (d) Identity Manipulation.

Fig. 5. Average spectrum images (SI) of four categories of fake faces (bottom) with different manipulation techniques and corresponding real faces (top). (a) Entire Synthesis. (b) Expression Manipulation. (c) Attribute Manipulation. (d) Identity Manipulation.
In detecting AI-manipulated fake faces, although existing detection methods consider different categories of fake faces, the detection performance on unseen manipulation techniques is still poor. The main reason is that on the one hand, the intrinsic features of fake faces are not well captured, on the other hand, there are differences among cross-manipulation techniques in feature distribution. Therefore, besides the previous module for mining intrinsic features, we also need to eliminate the distribution bias. In view of this, we propose a domain alignment module by minimizing the divergence among cross-manipulation techniques in feature distribution to reduce the distribution bias. The basic idea is that we enforce the domain alignment module to learn generalized features to further improve the generalization ability, as shown in Fig. 3 (c).

Based on the analysis above, we first partition faces into different domains according to different manipulation techniques. Specifically, the face samples with \( K \) seen domains for training are denoted as:

\[
X_d = [x_{d1}, \ldots, x_{dn_d}], \tag{7}
\]

where \( d \in \{1, \ldots, K\} \) and \( n_d \) is the number of samples of the domain \( d \). The corresponding labels are denoted as: \( Y_d = [y_{d1}, \ldots, y_{dn_d}] \) with 2 categories (fake/real). After mining intrinsic features, the input features of the \( m \)-th fully-connected layer is described as:

\[
V_F^m = [V_{F1}^m, \ldots, V_{FK}^m], \tag{8}
\]
**TABLE I**

| Dataset               | Manipulation Techniques | Collection        | Total Samples | Real Faces | Total Samples | Size      |
|-----------------------|-------------------------|-------------------|---------------|------------|---------------|-----------|
| Entire Synthesis      | StyleGAN                | Officially-released | 5000          | FFHQ       | 5000          | 1024 × 1024 |
|                       | StyleGAN2               |                   | 5000          |            | 5000          |           |
| Expression Manipulation | ExpertGAN             | Self-synthesis    | 5000          | CelebA     | 5000          | 128 × 128 |
|                       | GANimation              |                   | 5000          |            | 5000          |           |
|                       | HomoInterpGAN           |                   | 5000          |            | 5000          |           |
| Attribute Manipulation | CycleGAN                | Self-synthesis    | 5000          | CelebA     | 5000          | 128 × 128 |
|                       | StarGAN                 |                   | 5000          |            | 5000          |           |
|                       | STGAN                   |                   | 5000          |            | 5000          |           |
| Identity Manipulation | Faceswap                | FaceForensics++   | 15000         | FaceForensics++ | 15000      | 256 × 256 |
|                       | DeepFake                | DFDC              | 30000         | DFDC       | 30000         |           |

where $V_{F_d}^m \in \mathbb{R}^{n_d \times D}$ denotes the input features of the $m$-th fully connected layer from domain $d$. To eliminate the bias among cross-manipulation techniques in feature distribution, we propose a cross-domain alignment loss:

$$L_{CDA} = d \left( \hat{V}_F^m \right),$$

where $d()$ indicates the distance among cross-manipulation techniques. We use the Maximum Mean Discrepancy (MMD) measure in our work:

$$d \left( \hat{V}_F^m \right) = \frac{1}{K(K-1)} \sum_{d \neq j} \sum_{t_1=1}^{n_d} \sum_{t_2=1}^{n_j} V_{F_d,t_1}^m - \frac{1}{n_j} \sum_{t_2=1}^{n_j} V_{F_j,t_2}^m \right)^2$$

where $V_{F_d,t_1}^m \in \mathbb{R}^D$ denotes the input feature of $t$-th sample from $V_{F_d}^m$. Therefore, we train our framework from scratch on the face samples collected from multiple domains with cross-entropy loss $L_C$ and cross-domain alignment loss $L_{CDA}$. In the cross-domain alignment loss $L_{CDA}$, the MMD distance among the domains is required to be minimized. In other words, the network parameters can be learned as:

$$\Theta^* = \arg \min_\Theta L_C + \lambda L_{CDA},$$

where $\lambda$ is the weight of the cross-domain alignment loss.

Through explicitly restricting the bias among cross-manipulation techniques in feature distribution, our proposed framework is guided to learn domain-invariant representations to further improve generalization ability for detecting fake faces with unseen manipulation techniques.

**IV. EXPERIMENTAL RESULTS**

**A. Datasets**

The dataset used for experiments is constructed by four categories: face entire synthesis, facial expression manipulation, facial attribute manipulation and identity manipulation, and our dataset contains 10 state-of-the-art and popular manipulation techniques. For each dataset, we collect fake faces by generating them with the manipulation techniques or downloading the officially released generated faces. To make the distribution of the real and fake faces as close as possible, real faces are pre-processed according to the pipeline prescribed by each technique. Specifically, for the face entire synthesis dataset (with the size 1024 × 1024), 10000 real face images are collected from FFHQ [1], and 5000 synthesis faces are collected.
from two public datasets with StyleGAN and StyleGAN2, respectively. For the facial expression manipulation dataset (with the size $128 \times 128$), 15000 real faces are collected from CelebA [52], and corresponding 5000 manipulation faces generated by ExperGAN, GANimation and HomoInterpGAN, respectively. For the facial attribute manipulation dataset (with the size $128 \times 128$), 15000 real faces are collected from CelebA, and corresponding 5000 manipulation faces generated by CycleGAN, StarGAN and STGAN, respectively. For the identity manipulation dataset (with the size $256 \times 256$), we utilize the FaceSwap dataset of Faceforensics++ [22] and DeepFake dataset of DFDC [2]. Specifically, we first collect 1500 FaceSwap videos with different compression coefficient $c_0$, $c_23$ and $c_40$, 3000 DeepFake videos and the equal number of real videos. After that, we extract 10 frames from each video, and further use face detector MTCNN [53] to get the images of face region. If there are more than one face detected in a frame, only the largest one is extracted, and then the extracted faces are aligned to size of $256 \times 256$. Consequently, 15000 manipulation face images with FaceSwap, 30000 manipulation face images with DeepFake and the equal number of real face images are produced. Details of our dataset are summarized in Table I and examples of AI-manipulated fake faces are shown in Fig. 7.

### Table II

| Dataset                  | Face Entire Synthesis | Facial Expression Manipulation | Facial Attribute Manipulation | Identity Manipulation |
|--------------------------|-----------------------|-------------------------------|-------------------------------|-----------------------|
| Mesonet [20]             | 89.93                 | 90.95                         | 91.45                         | 89.99                 |
| Capsule [23]             | 91.78                 | 93.32                         | 92.21                         | 90.06                 |
| Xception [22]            | 93.05                 | 92.97                         | 92.05                         | 91.08                 |
| Color Spaces [43]        | 93.23                 | 93.3                          | 93.28                         | 91.38                 |
| Saturation Cues [39]     | 94.2                  | 93.42                         | 92.9                          | 91.34                 |
| Gram-Net [42]            | 94.05                 | 95.12                         | 96.45                         | 94.65                 |
| Self-Attention [41]      | 95.7                  | 95.55                         | 96.3                          | 94.65                 |
| FakeSpotter [27]         | 96.65                 | 96.88                         | 97.47                         | 95.68                 |
| $F^3$-Net [44]           | 96.83                 | 97.13                         | 97.77                         | 97.42                 |
| Proposed                 | 98.95                 | 99.57                         | 99.75                         | 98.97                 |

### Table III

| Manipulation Techniques | StyleGAN | StyleGAN2 | ExperGAN | GANimation | HomoInterpGAN | CycleGAN | StarGAN | STGAN | Faceswap | DeepFake |
|-------------------------|----------|-----------|----------|------------|---------------|----------|---------|-------|----------|----------|
| Mesonet [20]            | 90.2     | 89.66     | 90.4     | 90.65      | 91.8          | 90.55    | 91.15   | 92.65 | 90.02    | 89.98    |
| Capsule [23]            | 92.1     | 91.46     | 92.9     | 93.15      | 93.9          | 92.65    | 91.93   | 92.05 | 90.13    | 90.03    |
| Xception [22]           | 93.25    | 92.85     | 93.1     | 92.85      | 92.95         | 92.9     | 91.8    | 91.45 | 91.15    | 91.04    |
| Color Spaces [43]       | 93.55    | 92.9      | 92.95    | 93.1       | 93.85         | 93.95    | 92.85   | 93.05 | 91.77    | 91.18    |
| Saturation Cues [39]    | 94.95    | 93.45     | 93.8     | 93.25      | 93.2          | 92.95    | 92.9    | 92.85 | 91.48    | 91.27    |
| Gram-Net [42]           | 94.8     | 93.3      | 95.15    | 95.25      | 94.95         | 94.7     | 93.75   | 93.85 | 93.25    | 92.63    |
| Self-Attention [41]     | 96.1     | 95.3      | 96.15    | 95.4       | 95.1          | 96.55    | 96.3    | 96.75 | 94.25    | 94.85    |
| FakeSpotter [27]        | 96.9     | 96.4      | 96.5     | 96.75      | 97.4          | 97.15    | 97.3    | 97.95 | 96.72    | 95.16    |
| $F^3$-Net [44]          | 97.1     | 96.55     | 96.75    | 97.2       | 97.45         | 97.3     | 97.9    | 98.1  | 98.08    | 97.09    |
| Proposed                | 99.25    | 98.65     | 99.45    | 99.55      | 99.7          | 99.87    | 99.73   | 99.65 | 99.67    | 98.62    |

B. Implementation Details

The GPU card utilized for our task is NVIDIA GTX2080Ti and the framework is implemented by Torch library. The input size is $128 \times 128$, and we crop each image to several nonoverlapping $128 \times 128$ images. During training, we first train the framework only with cross-entropy loss, and then the last convolutional layer as well as the fully connected layer are fine-tuned with both cross-entropy loss and cross-domain alignment loss. Such training strategy is reasonable since shallow layers are more likely to be generalized [54] and more discriminative information expected to extract by fine-tuning.
TABLE IV
DETAILS OF CROSS-MANIPULATION TECHNIQUE EXPERIMENTAL PROTOCOLS.

| Protocol | Training Manipulated Techniques | Testing Manipulated Techniques | Protocol | Training Manipulated Techniques | Testing Manipulated Techniques |
|----------|---------------------------------|---------------------------------|----------|---------------------------------|---------------------------------|
| N1       | StyleGAN ExperGAN CycleGAN GANimation | StyleGAN2 | N5       | StyleGAN2 GANimation STGAN CycleGAN GANimation | StyleGAN |
| N2       | GANimation StarGAN DeepFake StyleGAN Faceswap | ExperGAN | N6       | StyleGAN2 HomOInterpGAN CycleGAN GANimation | GANimation |
| N3       | StyleGAN2 CycleGAN ExperGAN DeepFake STGAN | STGAN | N7       | DeepFake ExperGAN STGAN StyleGAN | StarGAN |
| N4       | Faceswap StyleGAN2 STGAN HomOInterpGAN CycleGAN | DeepFake | N8       | StyleGAN HomOInterpGAN DeepFake CycleGAN | Faceswap |

TABLE V
PERFORMANCE ON THE DETECTION OF FAKE FACES WITH UNSEEN MANIPULATION TECHNIQUES IN FOUR CATEGORIES OF DATASE (%).

| Protocol     | N1     | N2     | N3     | N4     | N5     | N6     | N7     | N8     |
|--------------|--------|--------|--------|--------|--------|--------|--------|--------|
| Zhang et al. | 90.82  | 91.39  | 91.17  | 90.74  | 90.19  | 90.72  | 90.57  | 90.07  |
| Li et al.    | 92.09  | 92.63  | 92.79  | 91.93  | 91.15  | 92.08  | 91.92  | 90.76  |
| Dang et al.  | 93.59  | 94.82  | 95.29  | 93.45  | 92.62  | 93.21  | 94.04  | 92.51  |
| Wang et al.  | 95.78  | 96.18  | 96.24  | 95.34  | 94.87  | 96.15  | 96.01  | 94.67  |
| Chai et al.  | 97.05  | 97.92  | 97.76  | 96.29  | 96.63  | 97.18  | 96.79  | 96.52  |
| Proposed     | 97.92  | 98.92  | 98.59  | 97.65  | 97.26  | 98.13  | 98.09  | 97.19  |

the deeper layers. For the learning parameter setting, we use the Adam optimizer \[55\] in a mini-batch manner with the size 10 during network initialization step and 100 for each domain during the fine-tuning step. The momentum values are set as \(\beta_1 = 0.9\) and \(\beta_2 = 0.999\), and the initial learning rate is set to \(1e^{-3}\) and \(1e^{-4}\) for fine-tuning. In addition, learning rates are dropped by \(10 \times \) if after 5 epochs the validation accuracy does not increase by 0.1\%, and we terminate training at learning rate \(1e^{-7}\). The weight \(\lambda\) of the cross-domain alignment loss is set in the way that at the end of training, the classification loss and cross-domain alignment loss are approximately the same. The reason for such setting is that our framework can learn both the discrimination and generalization ability. More specifically, the weight \(\lambda\) is selected in a range \([0.001, 0.01, 0.1, 1, 10]\). During testing, a query face is first cropped to several \(128 \times 128\) parts, from which we can recover the entire face image. Then, these cropped \(128 \times 128\) parts will be judged by independently. Finally, the face image is considered to be fake if one of the cropped parts is judged to be fake. In our experiments, 5 experiments are performed and the average result is reported each time.

C. Detection of Fake Faces with Seen Manipulation Techniques

In this section, we evaluate the performance of our framework on the detection of fake faces with seen manipulation techniques in all four datasets, that is, the manipulation technique in the testing sets also exists in the training sets. We merge four datasets together and the training, validation and testing sets are randomly selected. The ratio of face images in training, validation and testing sets is set to be 6:2:2 for each dataset. In order to better demonstrate the superiority of the proposed method, we compare the results with the state-of-the-art methods on detection of AI-manipulated fake faces. For comparison, their methods are applied to our datasets. The comparative results on four categories of fake faces are shown
in Table III. From the results, all the accuracies of our framework are higher than 98%, and outperform state-of-the-art methods in all cases. We further evaluate the performance of our framework on each manipulation techniques and compare the results with state-of-the-art methods, as shown in Table III. Experimental results demonstrate the excellent performance of our method, as well as a significant improvement compared with the state-of-the-art methods. Especially on the latest DFDC and StyleGAN2 datasets, we also achieve relatively high accuracies. This could be explained as below: compared to these methods, our framework not only exploits intrinsic clues from SI and CDI based on the fundamental differences between the AI-manipulation process and the camera imaging process, but also introduces the OctConv-based feature learning module and the attention-based feature fusion module to effectively mine intrinsic features for fake faces detection.

In summary, the proposed framework can effectively detect the AI-manipulated fake faces with seen manipulation techniques.

D. Detection of Fake Faces with Unseen Manipulation Techniques

In this section, to further evaluate the generalization capability of our proposed framework, we conduct experiments to evaluate the performance on the detection of fake faces with unseen manipulation techniques. Specifically, we first train a framework by employing fake faces produced by multiple manipulation techniques and corresponding real faces. Then we evaluate the performance by testing faces with another manipulation technique which is not involved in the training phase. To conduct such unseen-manipulation technique based experiments, we merge four datasets together and rearrange training, validation and testing sets based on manipulation techniques. In particular, we randomly create 8 different cross-manipulation technique scenarios to evaluate the performance of our framework, and the details of experimental protocols are demonstrated in Table V. Moreover, we compare the result with state-of-the-art methods focusing on generalization to demonstrate the generalization ability of our framework. For comparison, their methods are applied to our dataset. The comparative results are shown in Table V. It is observed that although the accuracies of detecting unseen manipulation techniques are slightly decreased than that of detecting seen manipulation techniques, but the results are all over 97%. Moreover, it can be obtained from the results that our proposed framework outperforms state-of-the-art methods in all cases, which proves that our framework has better generalization ability, and the generalization performance is better with more source domains: (Protocol N1-N4). This can be explained by the following: the universal detector or simulator may not be able to simulate the artifacts of all manipulation techniques in [30] and [31]. Moreover, compared to [32], [46] and [47], we mine more efficient intrinsic features and propose an alignment module to obtain generalized features for fake faces detection.

In a word, the proposed framework shows excellent advantages in improving generalization performance.

E. Ablation Study

In this section, we perform the ablation study to evaluate the performance of each component of our framework. We use the face entire synthesis and facial expression synthesis datasets, and the specific details of the dataset is shown as Table VI.

1) Results on Different Features: As described in Section III-B, our framework learn intrinsic features from CDI and SI instead of RGB images. To verify that the CDI and SI contain more intrinsic clues, we exploit the feature \( V_{RGB} \) learned from RGB images to detect fake faces on the dataset for ablation study. Moreover, the features \( V_{CDI} \) and \( V_{SI} \) are combined into \( V_F \) with the attention fusion module. Therefore, in order to analyze these two feature sets, we utilize \( V_{CDI} \) and \( V_{SI} \) to detect fake faces on the dataset used for ablation study. We further compare the simple concatenation fusion feature \( V_C \) and the attention fusion feature \( V_F \) to evaluate the performance of the attention fusion module. The results are shown in Table VII. From the results, the detection results based on \( V_{CDI} \) and \( V_{SI} \) achieve better performances than that on \( V_{RGB} \), due to the clues of CDI and SI are efficient and intrinsic. It is also observed that the results on the simple concatenation fusion feature \( V_C \) are all better than those on the feature set \( V_{CDI} \) or \( V_{SI} \). Moreover, the results on the attention fusion feature \( V_F \) are all best compared to those on other features, which demonstrate the necessity of attention feature fusion module.

2) Visualizations of Attention-based Feature Fusion Module: To explore the effectiveness of our attention-based feature fusion module, we further show the visualization results. Some samples are selected from four categories of fake faces datasets for analyzing the adaptive weighting mechanism of the fusion module. From the samples in Fig. 8, we can see the weights for CDI and SI are adaptively weighting. For the entire synthesis faces, the weights of SI are higher than those.
Entire Synthesis
Expression Manipulation
Attribute Manipulation
Identity Manipulation

Fig. 8. Attention fusion weights (numbers in the boxes) showing the importance of CDI and SI. Samples cover four categories of fake faces.

TABLE VIII
THE RESULTS OF ABLATION STUDY ON OCTCONV OPERATOR (%).

|                  | Without OctConv | With OctConv |
|------------------|-----------------|--------------|
| StyleGAN2        | 96.45           | 98.75        |
| HomoInterpGAN    | 96.75           | 99.65        |

of CDI because the clues from SI is more efficient. For the expression manipulation and attribute manipulation faces, the weights for CDI and SI are similar, and the weights of CDI are slightly higher than those of SI. For the identity manipulation faces, the CDI gain higher weights because of more pre- and post-processing in the video forgery process.

3) Importance of OctConv Operator: In order to evaluate the performance of the octconv operator, we test the framework with and without octconv operator on the dataset used for ablation study. The comparison results are shown in Table VIII. From the table, the framework with OctConv operator performs better, which prove the effectiveness of the OctConv Operator.

4) Importance and Visualizations of Domain Alignment Module: In this section, we further verify the effectiveness of domain alignment module in our framework. We test with and without the domain alignment module on the dataset used for ablation study. The comparison results are shown in Table IX. From the results, the domain alignment module increase the performance of the proposed framework, and it plays a key role in improving the generalization ability.

Moreover, we randomly select 200 samples of each manipulation technique from the dataset for ablation study and plot the t-SNE visualizations to analyze the feature space learned by introducing the domain alignment module, as shown in Fig. 9. It can be seen that the domain alignment module can make the features more robust in the feature space, which is more conducive to detect fake faces with unseen manipulation techniques.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel framework to effectively detect AI-manipulated fake faces, especially focus on how to mine the generalized features on detecting unseen manipulation techniques. To achieve the goal, we mine intrinsic features and further eliminate the distribution bias among cross-manipulation techniques. First, we mine two intrinsic clues from the CDI and SI, rather than depending on the specific defects in the manipulation process. Moreover, we adopt OctConv and an attention-based fusion module to mine two intrinsic features more effectively. Finally, to obtain a more generalized framework, an alignment module is proposed to reduce the bias among cross-manipulation techniques in feature distribution. The experimental results show that the performance of the proposed method outperforms the state-of-the-art works, especially the generalization performance on detecting unseen manipulation techniques. In the future work, we will consider more challenging situations and focus on the effectiveness in detecting AI-manipulated fake faces on social networks.

REFERENCES

[1] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability,
and variation,” in International Conference on Learning Representations, 2018.

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8110–8119.

[4] H. Ding, K. Sricharan, and R. Chellappa, “Exprgan: Facial expression editing with controllable expression intensity,” in Thirty-second AAAI conference on artificial intelligence, 2018.

[5] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “Ganimation: Anatomically-aware facial animation from a single image,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 818–833.

[6] Y.-C. Chen, X. Xu, Z. Tian, and J. Jia, “Homomorphic latent space interpolation for unpaired image-to-image translation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2408–2416.

[7] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2387–2395.

[8] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.

[9] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, “Attngan: Facial attribute editing by only changing what you want,” IEEE Transactions on Image Processing, vol. 28, no. 11, pp. 5464–5478, 2019.

[10] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, “Stgan: A unified selective transfer network for arbitrary image attribute editing,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 3673–3682.

[11] I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast face-swap using convolutional neural networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3677–3685.

[12] I. Petrov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, J. Jiang, L. RP, S. Zhang, P. Wu et al., “Deepfacelab: A simple, flexible and extensible face swapping framework,” arXiv preprint arXiv:2005.05335, 2020.

[13] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” ACM Transactions on Graphics (TOG), vol. 38, no. 4, pp. 1–12, 2019.

[14] S. Chen, S. Tan, B. Li, and J. Huang, “Automatic detection of object-based forgery in advanced video,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 11, pp. 2138–2151, 2015.

[15] L. M. Dang, S. I. Hassan, S. Im, and H. Moon, “Face image manipulation detection based on a convolutional neural network,” Expert Systems with Applications, vol. 129, pp. 156–168, 2019.

[16] X. Zhao, S. Wang, S. Li, and J. Li, “Passive image-splicing detection by a 2-d noncausal markov model,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 2, pp. 185–199, 2014.

[17] Y. Liu, X. Zhu, X. Zhao, and Y. Cao, “Adversarial learning for constrained image splicing detection and localization based on atrous convolution,” IEEE Transactions on Information Forensics and Security, vol. 14, no. 10, pp. 2551–2566, 2019.

[18] F. Matern, C. Rieß, and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” in 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019, pp. 83–92.

[19] Y. Li, M.-C. Chang, and S. Lyu, “In ictu oculi: Exposing ai created fake videos by detecting eye blinking,” in 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018, pp. 1–7.

[20] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” in 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018, pp. 1–7.

[21] N. Yu, L. S. Davis, and M. Fritz, “Attributing fake images to gans: Learning and analyzing gan fingerprints,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7556–7566.

[22] A. Rossler, D. Cozzolino, L. Verdoliva, C. Rieß, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1–11.

[23] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 2307–2311.

[24] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018, pp. 1–6.

[25] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 80–87.

[26] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, “Deepfake video detection through optical flow based cnn,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[27] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, “Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces,” arXiv, pp. arXiv–1909, 2019.

[28] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, “Incremental learning for the detection and classification of gan-generated images,” in 2019 IEEE International
Yang Yu received the B.S. degree from China University of Petroleum, Beijing, China, in 2017. She is pursuing the Ph.D. degree in Beijing Jiaotong University, China.

Rongrong Ni received the Ph.D. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2005. Since 2005, she has been the faculty of the School of Computer and Information Technology and the Institute of Information Science, where she was an Associate Professor from 2008 to 2013, and was promoted to a Professor in 2013.

Her current research interests include image processing, data hiding and digital forensics, pattern recognition, and computer vision.

Yao Zhao received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor at BJTU in 1998 and became a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding.