Face Forgery Detection via Symmetric Transformer

Luchuan Song*
University of Rochester
Rochester, New York, USA
lsong11@ur.rochester.edu

Xiaodan Li†
Alibaba Group
Hangzhou, Zhejiang, China
fiona.lxd@alibaba-inc.com

Zheng Fang*
Shopee Inc.
Shanghai, China
fangzheng0827@gmail.com

Zhenchao Jin
The University of Hong Kong
Hong Kong
blwx96@connect.hku.hk

YueFeng Chen
Alibaba Group
Hangzhou, Zhejiang, China
yuefeng.chenyf@alibaba-inc.com

Chenliang Xu
University of Rochester
Rochester, New York, USA
chenliang.xu@rochester.edu

ABSTRACT

The deep learning-based face forgery detection is a novel yet challenging task. Despite impressive results have been achieved, there are still some limitations in the existing methods. For example, the previous methods are hard to maintain consistent predictions for consecutive frames, even if all of those frames are actually forged. We propose a symmetric transformer for channel and spatial feature extraction, which is because the channel and spatial features of a robust forgery detector should be consistent in the temporal domain. The symmetric transformer adopt the newly-designed attention-based strategies for channel variance and spatial gradients as the vital features, which greatly improves the robustness of deepfake video detection. Moreover, this symmetric structure acts on temporal and spatial features respectively, which ensures the robustness of detection from two different aspects. Our symmetric transformer is an end-to-end optimized network. Experiments are conducted on various settings, the proposed methods achieve significantly improvement on prediction robustness and perform better than state-of-the-art methods on different datasets.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Symmetric transformer; Deepfake video detection

ACM Reference Format:
Luchuan Song*, Xiaodan Li†, Zheng Fang*, Zhenchao Jin, YueFeng Chen, and Chenliang Xu. 2022. Face Forgery Detection via Symmetric Transformer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547806

† Corresponding authors.
* Equal contribution.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547806

1 INTRODUCTION

Manipulated media (*e.g.*images and videos) with realistic forged faces can be easily generated by off-the-shelf softwares or algorithms, benefited from the rapid development of face forgery techniques [1, 2, 28, 45, 48, 49]. Although there are great prospects in the field of entertainment and human-computer interaction, these advanced forgery approaches may be abused for malicious purpose, causing trust issues and security concerns in public society. To take up these challenges, face forgery detection approaches have been explored based on not only the hand-crafted features [14, 20, 21, 40] but also deep learning representations [16, 18, 34, 50]. These approaches, especially the CNN-based methods, achieve remarkable performance boosts on several large-scale public datasets, *e.g.*FaceForensics++ [43] dataset.

However, with realistic and delicate faces manipulated, these "reliable" CNN models are found to hardly maintain robust and stable predictions on the images sampled from the same videos. As shown in Fig. 1(a), the widely used Xception [13] in face forgery detection gets conflicting scores on adjacent frames randomly captured from the same video. This phenomenon commonly exists in face forgery detection especially for those well-forged faces. Recently, there are several approaches [19, 35] that utilize 3D convolution or recurrent neural network to boost the performance of forgery video detection with exploration on the temporal-based representations of video frames. Although those methods may eliminate prediction instability by correlating multiple frames, the convolution or recurrent operations on the temporal dimension can hardly tackle the jitter problem in principle, and also bring high computational cost with 3D convolutions. The examples in Fig. 1(b) shows that the prediction instability of those videos still exists for SlowFast R-101 [19] (denoted as Slowfast).

To address the prediction instability on face forgery detection, we focus on the **robust** feature learning in our work. Inspired by the channel-spatial disentangling analysis on feature maps in the previous works [10, 22], we respectively explore the channel-wise and spatial-wise feature properties of hard cases. 1) *Channel-wise-* As shown in Fig. 1(c-1), the channel-wise variance of feature maps among those video frames distributes variously and white cells represent higher variance. The variance values may directly lead to the score instability of the final prediction and may further contain the potential clues for robust face forgery detection . In our work, we will explore how to bring the *channel variance* into the feature learning of current approaches. 2) *Spatial-wise-* The forgery

artifacts are usually contaminated by compression error or refined by the well-designed forgery methods. These "flickering" artifacts sometimes can hardly be accurately captured in general spatial domain (RGB). That is another main culprit for the score instability of face forgery detection, apart from the instable channel-wise distribution. As shown in Fig. 1(c-2), there still exists inconspicuous gradient clues to help obtain the effective features while the forgery artifacts are hardly extracted from the general feature maps. Therefore, we intend to directly highlight the gradient patterns at multiple feature levels to mine the potential artifacts of face forgery. Motivated by the aforementioned observations, we propose a novel end-to-end framework, named as *RobustForensics* to capitalize more efficient and robust feature learning for face forgery detection. To further incorporate the channel variance and spatial gradient into long-range dependencies modeling which are emphasized in the recent transformer-based methods [5, 9, 17, 46, 47, 52, 60], we modulate our modules based on attention-based structure. The proposed framework contains two important modules,

1) *Variance-Restricted Channel Activation (VRCA)* focuses on modeling the feature correlation on channel-wise distribution among those "similar" frames and adaptively activates channels based on its variance to enhance feature learning through channel-wise attention.

2) *Gradient-Enhanced Spatial Attention (GESA)* extracts potential patterns by calculating the spatial neighboring gradients at multiple feature levels. Our gradient enhancement policy is combined with attention block and further strengthen the characteristic of potential feature mining.

The proposed VRCA and GESA are utilized to enhance the feature learning from the complementary dimensions, *i.e.*channel-wise and spatial-wise respectively, and then merged with original input feature by element-wise sum as output in each stage. We conduct experiments from the perception of in-domain performance (FaceForensics++ (FF++) [43]), generation to distortion perturbations (DeeperForensics [28]) and generalization to unseen data (Celeb-DF v2 [32]) on widely-used datasets, our RobustForensics gains significant improvements compared with baseline and better than state-of-the-art approaches.

2 RELATED WORK

Forgery Detection. Early approaches mainly focus on examining the appearance features on the spatial domain such as RGB and HSV. Some studies [16, 23, 29, 37] extract color-space features for classification. Recent methods [8, 16, 30, 33, 50] use deep neural networks to extract high-level information from the spatial domain and result in considerable improvements. GramNet [33] extracts global textures to tackle the distortion perturbations. Face X-ray [30] explores the detection task on locating the boundary of face forgery. Recently there is a growing number of studies [11, 35, 41, 53, 55] pay attention on frequency features. F^3-Net [41] uses frequency-aware image decomposition and local frequency statistics to mine forgery patterns while Two-branch RNN [35] uses the Laplacian of Gaussian operator to merge information extracted from both color domain and the frequency domain. All these approaches only focus on exploring more forgery information, without considering the robustness of extracted feature. Our proposed RobustForensics explores the unrobust channel based on channel-wise variance and highlights the potential clues through spatial gradient enhancement.

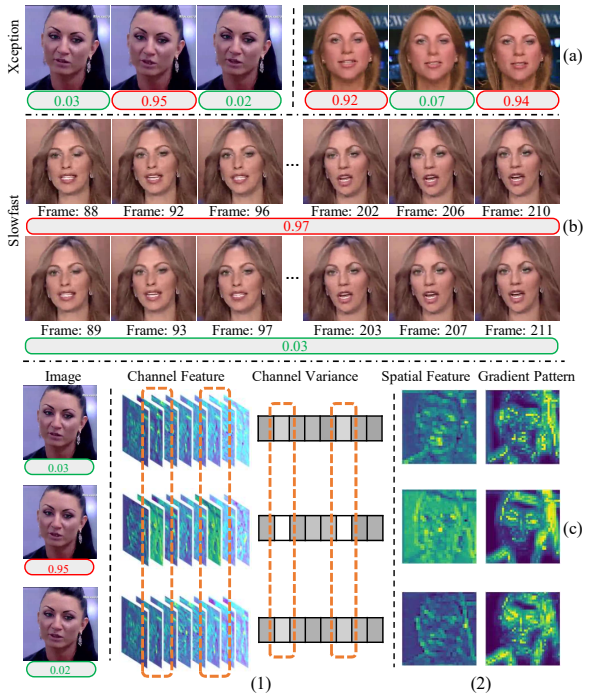

Figure 1: The frames from the same video get totally different predictions on both image-based approach in (a) and video-based approach in (b). Higher scores (red) denote that prediction results are fake and lower scores (green) for real. The index of corresponding frame is listed under the image in (b). Even with only one frame shift, the prediction score gap is big. The channel variance and the spatial gradient are shown respectively in (c). Channels with higher variance are highlighted in the dashed box. White cells represent higher variance in (c-1) while gradient patterns are extracted by spatial features in (c-2).

Attention Mechanism. The attention mechanism in deep learning gets more popular in the last few years. Transformer [52] is proposed for natural language processing (NLP) which is based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Bottleneck Transformer [46] is proposed to perform transformer encoder under the bottleneck structure [25]. Both Non-local network [56] and Bottleneck Transformer [46] are spatial-wise attention blocks. SE-Net [27] focuses on the channel-wise attention to explicitly model channel interdependencies. Our proposed RobustForensics involves both spatial gradient and channel variance into traditional attention mechanism from spatial-wise and channel-wise dimension respectively to enhance the potential feature mining for face forgery detection.

3 ROBUSTFORENSICS

In our work, we introduce a novel end-to-end architecture named as *RobustForensics* to boost the prediction robustness from two aspects: channel-variance activation and spatial-gradient enhancement. As shown in Fig. 2, several frames are utilized as input of the

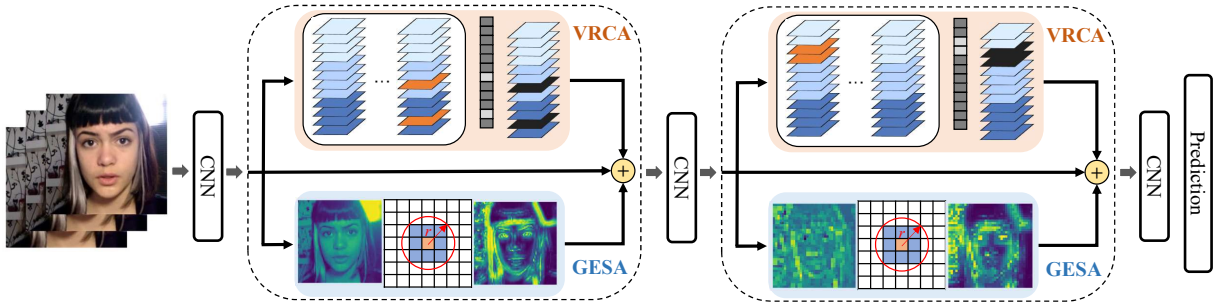

Figure 2: The architecture of RobustForensics. In the aspect of channel, VRCA utilizes the feature variance of each channel among frames to adaptively activate channels. As for spatial, gradient patterns are augmented through our GESA. The refined features by VRCA and GESA are merged with original feature as output of the current stage.

whole framework. To effectively activate channel-wise feature, we propose a *Variance-Restricted Channel Activation (VRCA)*, which subtly utilizes the feature variance of each channel among multiple frames to enhance the robustness of learned patterns. As for the spatial aspect, gradient patterns are augmented through our *Gradient-Enhanced Spatial Attention (GESA)*, which highlights the context gradient for those potential artifacts. Then, the refined features by VRCA and GESA are merged with original feature as output of the current stage. To further incorporate the channel variance and spatial gradient into long-range dependencies modeling, we modulate the feature learning based on attention-based structures. Worth noting that benefited from the modular design of our proposed module, it can be easily used in a plug-and-play manner at different stages of various backbone structures, not only for the attention blocks.

3.1 Variance-Restricted Channel Activation

Channel-wise attention is widely used in recent approaches [10, 12, 22, 27] which explicitly models channel interdependencies. However, the channel-wise instability on encoding patterns still exists in these works, especially for face forgery detection. In general, the face forgery is performed for a certain period of time and the most of adjacent frames are supposed to contain the similar feature. Motivated by this characteristic, we adaptively activate channel-wise features based on the variance among multiple frames, defined as *Variance-Restricted Channel Activation* (VRCA). The architecture of primary channel-wise attention is modified to modulate compatibility weights based on channel variance.

Channel-wise Attention. Recent approaches [10, 22] utilize a transformer-based structure on channels to correlate their dependencies with rich context. The attention function is described as mapping a query and a set of key-value pairs to an output, where queries ($\mathbf{Q}$), keys ($\mathbf{K}$), values ($\mathbf{V}$) are extracted from the input feature by three 1×1 convolution layers: $\mathbf{W}_\mathrm{Q}$, $\mathbf{W}_\mathrm{K}$ and $\mathbf{W}_\mathrm{V}$ respectively. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. Formally,

$$\mathbf{F}^\mathrm{C} = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{HW}})\mathbf{V}, \tag{1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{C \times HW}$ and C, H, W corresponds to channel, height and width respectively. The obtained channel attention map

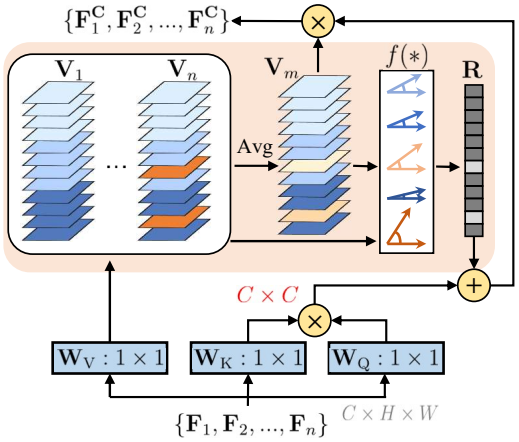

Figure 3: The architecture of VRCA is constructed based on channel-wise attention. Channel variance among multiple frames is extracted from $\{\mathbf{V}\}$ and the attention map is modulated based on channel variance to consider both interdependency and stability. Multi-Head Attention is not present for better visualisation.

encodes the interdependencies between channels and the output consists the long-range pattern correlations between feature maps.

Variance-Restricted Channel Activation. To model the relationship of the adjacent frames in the channel-wise attention module, the value variance is calculated on $\{\mathbf{V}\}$. In our VRCA, as shown in Fig. 3, multi-frame values set $\{\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_n\}$ is extracted from input features $\{\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_n\}$, where n is the number of frames. Then, $\mathbf{V}_m$ is the mean of $\{\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_n\}$. Specifically, $\mathbf{V}_m = \frac{1}{n}\sum_i^n \mathbf{V}_i$, where i represents the i-th frame. The distance between $\mathbf{V}_m$ and $\mathbf{V}_i$ under the same channel index reflects the the robustness of current channel on encoding certain patterns. Cosine distance has the advantage of measuring the similarity between two distributions compared with Euclidean distance and is utilized in VRCA. Formally, the cosine distance f is utilized to measure the channel-wise difference between $\mathbf{V}_m$ and $\mathbf{V}_i$,

$$f_c(\mathbf{V}_m, \mathbf{V}_i) = 1 - \frac{\mathbf{V}_m(c)\mathbf{V}_i^\top(c)}{\|\mathbf{V}_m(c)\|_2 \|\mathbf{V}_i(c)\|_2}, \tag{2}$$

where c means the index of channel of the feature maps and $f_c(*)$ represents the function on the c-th channel. The channel-wise variance is obtained through the average distance of each frame

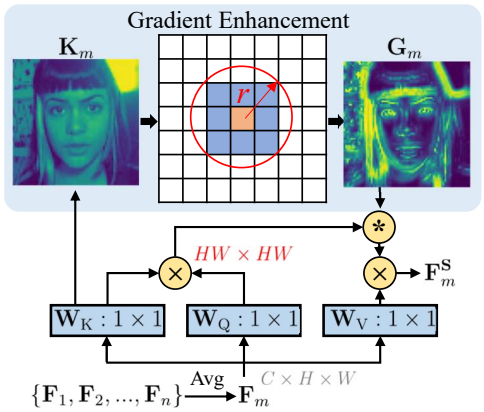

Figure 4: The architecture of GESA is constructed based on spatial-wise attention. Gradient patterns are extracted from $\mathbf{K}_m$ through the spatial neighbouring gradients and are involved in feature learning by taking the gradient patterns as an attention mask to highlight information-rich area. Multi-Head Attention and the position encoding are not present for better visualisation.

value $\mathbf{V}_i$ and the corresponding mean value $\mathbf{V}_m$. To a certain extent, the feature stability are negatively related to the variance values. Formally,

$$\mathbf{R} = 1 - \frac{1}{n}\sum_{i}^{n} f(\mathbf{V}_m, \mathbf{V}_i), \tag{3}$$

where $\mathbf{R} \in \mathbb{R}^{1\times C}$ means the quantified coefficient based on channel-wise variance. The value range of $\mathbf{R}$ is $[-1, 1]$. To enhance feature learning during the end-to-end training, $\mathbf{R}$ is added row by row in the attention map through an adaptive weight λ. Furthermore, the mean value $\mathbf{V}_m$ is utilized to replace the instance value $\mathbf{V}_i$. Therefore, the traditional channel attention is modified as follows,

$$\mathbf{F}^{\mathrm{C}}_i = \mathrm{softmax}(\frac{\mathbf{Q}_i\mathbf{K}_i^{\top}}{\sqrt{HW}} + \lambda\mathbf{R})\mathbf{V}_m. \tag{4}$$

Worth noting that, compared with primary channel-wise attention block as shown in Eq. 1, few additional parameters are involved in the proposed VRCA and the channel-wise variance among multiple frames are introduced to enhance feature learning during the end-to-end training.

3.2 Gradient-Enhanced Spatial Attention

To further mine the potential artifacts caused by compression error or subtle forgery, we focus on the spatial-wise feature learning, regarded as the complementary operation compared with aforementioned channel-wise attention.

In our work, we innovatively involve the gradient patterns in the feature learning procedure from a particular perspective by taking the spatial gradients as attention to enhance spatial feature learning. The proposed module is named as *Gradient-Enhanced Spatial Attention* (GESA). As shown in Fig. 4, the architecture of GESA is based on attention-based structure [52]. The primary function of the query with the corresponding key is calculated along the spatial-wise dimension formatted as follows.

$$\mathbf{F}^{\mathrm{S}} = (\mathrm{softmax}(\frac{\mathbf{Q}^{\top}\mathbf{K}}{\sqrt{C}}))\mathbf{V}, \tag{5}$$

To extract robust gradient information, the mean of $\{\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_n\}$ is utilized as input instead of the instance feature to weaken the noise interference. Specifically, $\mathbf{F}_m = \frac{1}{n}\sum_i^n \mathbf{F}_i$. Queries ($\mathbf{Q}_m$), keys ($\mathbf{K}_m$) and values ($\mathbf{V}_m$) are extracted from $\mathbf{F}_m$ by 1×1 convolution layer respectively. In the primary transformer structure, the attention map is calculated based on the compatibility between the query with the corresponding key. However, as mentioned in Informer [61], similar values in attention map will decrease the diversity of extracted feature. To this end, the proposed GESA performs gradient enhancement on the $\mathbf{K}_m$ to highlight the important keys in potential regions, which subtly merges gradient patterns to the attention map. In the procedure of gradient enhancement on feature maps, inspired by Sobel operator [44] in gradient calculation, the neighbouring feature is utilized to calculate the gradient through a sliding window with fixed weights. Formally, for the gradient-enhanced map $\mathbf{G}_m$, $\mathcal{S}_r(x)$ is used to represent the neighbour feature set where $x \in [0, HW)$ represents the location of current feature and r is the radius to select neighbourhoods ($r = 3$ in our experiments). Specifically,

$$\mathbf{G}_m(x) = 2 * \sigma(\frac{1}{|\mathcal{S}_r(x)|}\sum_{y\in\mathcal{S}_r(x)} \|\mathbf{K}_m(x) - \mathbf{K}_m(y)\|_2^2) - 1, \tag{6}$$

where $|.|$ means the number of elements in set and σ is the sigmoid function to normalize the value to $(0, 1)$. Due to the sigmoid of positive values is in the range of $[0.5, 1)$, a simple linear transformation is utilized in Eq. 6 to normalize the value in $\mathbf{G}_m$ between 0 and 1. As shown in Fig. 5(c), in $\mathbf{G}_m$, gradient patterns are highlighted by the mask in the attention map to focus on information-rich areas. $\mathbf{G}_m$ is multiplied to each row of traditional attention map. Then, the final output of the proposed GESA is obtained by matrix multiplication, formatted as follows.

$$\mathbf{F}^{\mathrm{S}}_m = (\mathrm{softmax}(\mathbf{G}_m * \frac{\mathbf{Q}_m^{\top}\mathbf{K}_m}{\sqrt{C}}))\mathbf{V}_m. \tag{7}$$

Both gradient enhancement and attention sparsity are considered in $\mathbf{F}^{\mathrm{S}}_m$ which helps the network to pay more attention on the invisible gradient information and improves the prediction robustness. Meanwhile, the proposed GESA involves few additional parameters during the end-to-end training and subtly merge the gradient patterns through an attention based block.

In order to take full advantage of features refined by channel variance and spatial gradient, we aggregate features from these two attention modules. Specifically, following the architecture of Bottleneck Transformer [46], VRCA and GESA are integrated under the bottleneck structure. Then, the refined features by VRCA and GESA are merged with original input feature by element-wise sum as output in each stage. Noted that our attention modules are simple and can be directly inserted in the existing detection pipeline. They do not increase too many parameters yet strengthen feature learning effectively.

Table 1: Quantitative results on FF++ dataset with all quality settings, *i.e.*LQ indicates low quality (heavy compression), HQ indicates high quality (light compression) and RAW indicates raw videos without compression. The bold results are the best. The reported approaches are spited based on whether utilizing 3D Convolution in backbone. The results of Xception are obtained from our experiments for fair comparison. The Acc score of F^3-Net [41] with threshold of 0.5 is copied from the supplementary material of F^3-Net.

Methods	AUC (LQ)	Acc (LQ)	AUC (HQ)	Acc (HQ)	AUC (RAW)	Acc (RAW)
Steg.Features [21]	-	55.98%	-	70.97%	-	97.63%
LD-CNN [15]	-	58.69%	-	78.45%	-	98.57%
Constrained Conv [4]	-	66.84%	-	82.97%	-	98.74%
CustomPooling CNN [42]	-	61.18%	-	79.08%	-	97.03%
MesoNet [3]	-	70.47%	-	83.10%	-	95.23%
Face X-ray [30]	0.616	-	0.874	-	-	-
Two-branch RNN [35]	0.911	86.34%	0.991	96.43%	-	-
Xception ** [13]	0.917	84.02%	0.963	95.04%	0.992	98.77%
Bottleneck Transformers [46]	0.926	85.27%	0.971	96.23%	0.995	98.93%
F^3-Net (Xception) * [41]	0.933	86.89%	0.981	97.31%	0.998	99.84%
RobustForensics (Xception)	**0.951**	**89.91%**	**0.999**	**99.11%**	**1.000**	**99.91%**
3D ResNet [24]	-	83.86%	-	-	-	-
3D ResNeXt [57]	-	85.14%	-	-	-	-
I3D [7]	-	87.43%	-	-	-	-
Slowfast [19]	0.936	88.25%	0.982	96.92%	0.994	99.34%
F^3-Net (Slowfast) * [41]	0.958	92.37%	0.993	98.64%	0.999	99.91%
RobustForensics (Slowfast)	**0.986**	**95.45%**	**1.000**	**99.20%**	**1.000**	**100.00%**

4 EXPERIMENTS

4.1 Setting

Datasets. We conduct experiments on widely-used datasets, *i.e.*Celeb-DF v2 [32], DeeperForensics [28] and FaceForensics++ (FF++) [43]. We follow previous settings used in their corresponding datasets and compare with other methods respectively. More details on these datasets are described in following. 1) FaceForensics++ (FF++) is a face forgery detection video dataset containing 1,000 real videos, in which 720 videos are used for training, 140 videos are reserved for validation and 140 videos for testing. Most videos contain frontal faces without occlusions and were collected from Youtube with the consent of the subjects. Each video undergoes four manipulation methods to generate four fake videos, therefore there are 5,000 videos in total. When training and evaluating on FF++, we follow the sampling strategy mentioned in [43] that samples 270 frames per video for the training and 100 frames per video for validation and testing. We evaluated all no compression (raw), medium compression (c23) and high compression levels (c40) subsets. 2) DeeperForensics is a newly proposed face forgery dataset containing 1,000 real videos the same with FF++ c23 real videos and 1,000 fake videos generated using the Variational Auto-Encoder proposed in [28]. The training, validation and testing are separated different from FF++ with 703 videos for training, 96 videos for validation and 201 videos for testing. DeeperForensics performs different level distortion perturbations on data and level-5 is the hardest level for detection. Following the setting described in [28], we use the hardest setting that training on raw data without distortion perturbations and testing on both level-5 and random-level data to validate the generalization of our method to distortion perturbations. 3) Celeb-DF v2 contains 5,639 fake videos and 590 real videos. Following the previous setting in [32, 35], we use the Celeb-DF v2 dataset to evaluate the generalization performance of our model on unseen data. We use the model trained on FF++ c40 to evaluated on Celeb-DF v2 test set with 518 videos.

Metrics. We apply the *Area Under the Receiver Operating Characteristic Curve (AUC)* and *Accuracy (Acc)* as our evaluation metrics on general face forgery detection following previous methods [30, 35, 41]. In our experiments, AUC is used as the main metric since it is not affected by class imbalance and threshold. Although the Acc is also widely-used in face forgery detection, we argue that the Acc is improper for this task, mainly caused by the sensitivity on class imbalance and the choice of threshold as mentioned in [35]. For fair comparison, the Acc is calculated with the threshold of 0.5 without any threshold adjusting tricks. In datasets where the Acc is reported (*i.e.*, FF++ and DeeperForensics), we follow the rules released by the official [28, 43] and use the real-fake balanced accuracy.

To quantify stability, two metrics are utilized in experiments: *Proportion of Unstable Predictions (PUP)* and *Correction Rate (CR)*. PUP_θ represents the proportion of unstable videos with the max score gap among frames higher than θ, where smaller PUP_θ means better stability. CR represents the proportion of same-video frame pairs which are originally get different predictions in baseline and are corrected after applying other methods, where higher CR corresponds to better ability on improving stability.

Implementation Details. In our experiments, Xception [13] pretrained on the ImageNet dataset is used as backbone. The newly-introduced layers and blocks are initialized with the Kaiming initialization [26]. Data augmentations like cutout, random crop, random flip, *etc.*are used when training. The input images are resized to

299. The networks are optimized via SGD with base learning rate as 0.2 and multi-step learning rate scheduler. The momentum is 0.9, and the batch size is set to 128.

There are also several previous studies [41, 57] which use 3D convolution as input for modeling temporal relations on face forgery detection. To demonstrate the generalization ability of proposed methods, we also plant our major modules, *i.e.*VRCA and GESA into SlowFast R-101 [19] (denoted as Slowfast briefly) pre-trained on Kinetics-700 [6]. The Slowfast-based RobustForensics uses the frames in the slow pathway of Slowfast. The same data augmentations mentioned above are used when training. The input images are resized to 224. The networks are also optimized via SGD and the base learning rate is set as 0.02. The momentum is 0.9, and the batch size is set to 64.

4.2 Comparison with previous methods

In this section, we compare our method with previous face forgery detection methods on FF++, DeeperForensics and Celeb-DF v2 datasets.

Face Forgery Detection. The results of the most-widely-used FF++ dataset are listed in Tab. 1. Our method RobustForensics outperforms all the previous methods on all quality settings, *i.e.*, LQ (c40, compressed with the quantization parameter equal to 40), HQ (c23, quantization of 23) and RAW respectively. The reported approaches are split based on whether utilizing 3D Convolution in backbone. All the Acc score of our approaches is obtained with the constant threshold of 0.5, without any threshold searching. For fair comparison, the Acc score of F^3-Net [41] with threshold of 0.5 is copied from the supplementary material of F^3-Net. Benefited from eliminating the weakness of prediction instability, our Xception-based model performs much better than other image-based approaches, with 0.951 in AUC score and 89.91% in Acc score respectively, which is even better than most video-based approaches (except the F^3-Net (Slowfast) [41]). When utilizing the same backbone (*i.e.*, Slowfast [19]), our RobustForensics gains significant improvement compared with F^3-Net (Slowfast), with 0.986 and 95.45% of AUC and Acc scores in comparison to 0.958 and 92.37%, in LQ task. Our RobustForensics gains stable improvement on both image-based backbone and video-based backbone, indicating that the problem of prediction instability is widely existed in current approaches and RobustForensics can effectively boost performance from the complementary dimensions.

Generalization on Distortions. To evaluate the generalization ability of the proposed methods to distortion perturbations, we conduct experiments on DeeperForensics [28] as there are abundant distortions including Gaussian blur, JPEG compression and *etc.*. The general setting [28] of this dataset is that training on raw data without distortion perturbations and testing on both level-5 (the highest distortion level) distortions and random-level distortions data. The performance of previous approaches reported in [28] and our method are listed in Tab. 2 (The results of previous works are copied from DeeperForensics [28] for a fair comparison and only Acc is reported in [28]). The method number in Tab. 2 is smaller than Tab. 1 as the DeeperForensics [28] is a newly published dataset and fewer methods are reported.

Worth noting that the naive Xception preforms bad on distortion perturbations with 88.38% Acc score on level-5 distortions and 94.75% on random-level distortions. The proposed channel-wise robust activation and spatial-wise gradient enhancement spur our Xception-based RobustForensics to achieve 93.78% Acc score on level-5 distortions and 97.76% Acc score on random-level distortions, even better than video-based approaches. When equipping our VRCA and GESA on Slowfast, the accuracy is further improved with 97.26% Acc score on level-5 distortions and 99.00% Acc score on random-level distortions, which demonstrates that our proposed module can further improve the performance on the basis of temporal modules.

Table 2: The binary detection accuracy when trained and tested on DeeperForensics-1.0 dataset with different distortion perturbations. The model is trained on the standard set without distortions (std), and tested on the standard set with single-level distortions (std/sing) and with random-level distortions (std/rand). The reported approaches are spited based on whether utilizing 3D convolution in backbones or not.

Methods	Acc (std/sing)	Acc (std/rand)
Xception [13]	88.38%	94.75%
ResNet+LSTM [28]	90.63%	97.13%
TSN [54]	91.50%	95.00%
RobustForensics (Xception)	93.78%	97.76%
C3D [51]	87.63%	92.38%
I3D [7]	90.75%	96.88%
Slowfast [19]	91.79%	96.52%
RobustForensics (Slowfast)	**97.26%**	**99.00%**

Generalization on Cross-dataset. We perform generalization experiments on Celeb-DF v2 [32] with the models trained on FF++ c40 (low quality). The results are listed in Tab.3. Following previous settings in [32, 35], frame-level AUC scores on FF++ Deepfakes subset and Celeb-DF v2 set is presented in experiments. The frame-level AUC of the state-of-the-art approach Two-branch RNN [35] is copied from the original paper [35] and results of other approaches are copied from Celeb-DF v2 [32]. Constrained by frame-level results reported in Celeb-DF v2 [32], only Xception-based RobustForensics is presented in Tab.3.

Our RobustForensics performs better than other approaches while maintaining the performance of FF++ Deepfakes. This performance gap is obvious when comparing with baseline that RobustForensics improves the AUC score of Celeb-DF v2 from 0.655 to 0.790. Benefited from the robust feature activation in VRCA, over-fitting features are inactivated and the communal patterns existed in various datasets are enhanced, which is critical for generalization. Besides, the gradient enhancement in GESA helps the model to explore subtle artifacts, which enables RobustForensics to adapt to various forgery methods.

Comparison on Attention-based Approaches. Both our VRCA and GESA are integrated to the model through the attention-based structure. Recently, various attention-based approaches are proposed including SAN [59] which utilizes self-attention mechanism to refine the feature, and ViT [17] which replaces the convolutional layers with self-attention and cross-attention structures. We

Table 3: Frame-level AUC on FF++ Deepfakes and Celeb-DF v2. Our RobustForensics is trained on FF++ c40 dataset.

Methods	FF++ [DF]	Celeb-DF v2
Two-stream [62]	0.701	0.538
Meso4 [3]	0.847	0.548
MesoInception4 [3]	0.830	0.536
HeadPose [58]	0.473	0.546
FWA [31]	0.801	0.569
VA-MLP [36]	0.664	0.550
VA-LogReg [36]	0.780	0.551
Xception-c40 [13]	0.955	0.655
Multi-task [38]	0.763	0.543
Capsule [39]	0.966	0.575
DSP-FWA [31]	0.930	0.646
Two-branch RNN [35]	0.932	0.734
RobustForensics (Xception)	**0.971**	**0.790**

Table 4: Performance of different attention-based approaches on FF++ c40.

ID	Methods	AUC	$PUP_{0.5}$
1	Xception (Baseline)	0.917	65.9%
2	SAN-19 [59]	0.922	68.7%
3	ViT-B16 [17]	0.910	49.6%
4	Ours (Xception)	**0.951**	**34.4%**

also make comparisons with recent proposed attention-based approaches, as shown in Tab. 4. Our proposed RobustForensics performs better than other attention-based approaches in both AUC score and stability, which demonstrates the advantages of channel-wise variance and spatial-wise gradient.

Although there is performance improvement on SAN [59] compared with baseline, the stability becomes even worse which may be caused by missing position information in the patch-wise self-attention. The transformer-based ViT [17] performs bad on AUC while gains a significant improvement on stability. This phenomenon may be caused by the large number of parameters in ViT where the model tends to be over-fitted on the training set and generalizes badly when testing. Different from SAN, position embedding is incorporated in ViT which may be the reason of better stability. Although the ViT performs better than baseline on stability, the final prediction performance is worse. By contrast, our RobustForensics performs better in both AUC and stability, which is critical in practical.

4.3 Ablation Study

The ablation study experiments are performed on FF++ c40 dataset due to it is challenging to most of previous methods. The Xception-based models are used since it is lightweight and is widely-used for forgery detection in previous works. The AUC score is used as the metric.

Effectiveness of VRCA & GESA. To evaluate the effectiveness of the proposed VRCA and GESA, we quantitatively evaluate our model and its variants: 1) the naked Xception as the baseline, 2) Xception with VRCA, 3) Xception with GESA, 4) Xception with both VRCA and GESA (RobustForensics). Both the AUC score and the prediction stability are reported in Tab. 5. Smaller PUP_θ represents better stability with a small prediction score gap among frames. Higher CR corresponds to better ability on improving stability.

Table 5: Ablation study of our method on FF++ c40 (low quality). PUP_θ represents the proportion of videos with the frame-level score gap higher than θ, the smaller the better. CR represents the proportion of corrected unstable frame pairs in baseline, the bigger the better.

ID	VRCA	GESA	AUC	$PUP_{0.7}$	$PUP_{0.5}$	$PUP_{0.3}$	CR
1	-	-	0.917	60.7%	65.9%	71.6%	-
2	√	-	0.939	26.1%	35.6%	43.7%	82.9%
3	-	√	0.940	31.3%	37.3%	45.4%	76.1%
4	√	√	**0.951**	**24.1%**	**34.4%**	**42.0%**	**86.0%**

As shown in Tab. 5, even if only utilizing VRCA (model 2) or GESA (model 3), significant improvement on detection performance is achieved with AUC score 0.939 and 0.940 respectively. When using both VRCA and GESA (model 4), our method achieves the best performance with 0.951 AUC, much better than the 0.917 of baseline. As shown in ROC curves in Fig. 5(a), RobustForensics receives the best performance at lower false positive rate (FPR), while low FPR rate is a most challenging scenario with high restriction on prediction stability. The results of stability are positively related with detection performance where higher AUC corresponds to lower PUP and higher CR. The $PUP_{0.7}$ of RobustForensics is smaller than half of the baseline with 24.1% in comparison to 60.7%, and 86.0% inconsistent frame pairs in baseline are corrected by RobustForensics.

Furthermore, As shown in Tab.6, to demonstrate the improvement in RobustForensics is not introduced by simply using multiple frames, we conduct experiments on multiple frames with other components: simply stacking multiple frames (model 2), 3D Conv (model 3) and primary structure with channel&spatial attention (model 4). Our proposed RobustForensics with VRCA and GESA performs better than these above methods. Worth noting that, the 3D convolution based model (model 3) utilizes the similar architecture with RobustForensics by replacing the VRCA and GESA with 3D convolution, and the number of parameters in 3D convolution is about 4.5 times of RobustForensics when the frame number is 3. When only utilizing the basic channel and spatial attention architecture (model 4), the improvement on both performance and stability is limited compared with our RobustForensics (model 5). The improvement of stability comes from our proposed two components, especially the VRCA, where the activation on robust patterns and the enhancement on gradient patterns play a vital role on boosting stability. Furthermore, to demonstrate the complementarity of our method with 3D convolution on boosting stability, we conduct experiments on Slowfast by shifting the start frame index with [0, 9] frames. PUP is calculated on each video under the 10 predictions and the results are presented in Fig. 5(b). With the help of VRCA and GESA, the PUP of RobustForensics is only about half of Slowfast baseline on various score gap thresholds, demonstrating the complementarity of RobustForensics with general 3D convolution operation.

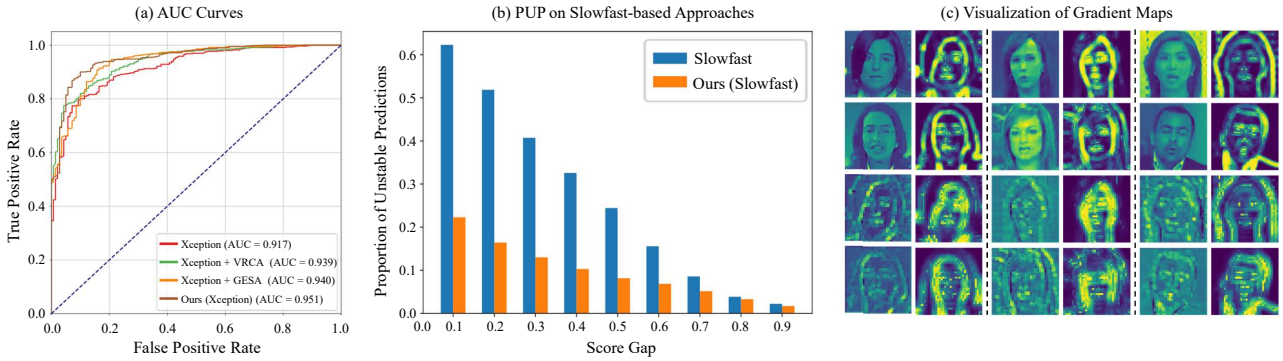

Figure 5: (a) ROC Curves of models in our ablation studies. (b) Comparison on prediction stability between Slowfast baseline and our Slowfast-based RobustForensics. Smaller PUP means better stability. (c) The visualization of gradient maps extracted by GESA. Every two columns represent the original feature (left) and the extracted gradient map (right) respectively. First two rows show the feature in early stages and last two rows correspond to later stages.

VRCA. To demonstrate the benefits of utilizing channel variance in VRCA, we evaluate the proposed VRCA and its variants by removing or replacing some components, *i.e.*, 1) Xception (baseline), 2) the proposed VRCA with only channel variance, denoted as "+ Channel Variance", 3) the proposed VRCA with only channel attention, denoted as "+ Channel Attention". All the experiments are under the same hyper-parameters for fair comparisons.

The performance of each variants is listed in Tab. 6. The improvement of VRCA (model 8) is mainly from the channel variance (model 7), and channel-wise attention (model 6) further brings the variance information into feature learning. When only applying the channel attention [22] or channel variance on Xception one by one, the performance will also achieve an obvious gain comparing with the baseline (model 1). By comparing the results of model 6 and model 7, it shows that the improvement of VRCA benefits more from the information complement of channel-wise variance.

GESA. To demonstrate the effectiveness of gradient enhancement in our proposed GESA, we conduct experiments on the variants of GESA by removing or replacing some components, *i.e.*, 1) Xception (baseline), 2) proposed GESA with only gradient enhancement, denoted as "+ Gradient Enhancement", 3) proposed GESA with only spatial attention, denoted as "+ Spatial Attention", 4) Xception with the frequency components proposed by F^3-Net[41], denoted as " + Frequency". The results are listed in Tab. 6.

The improvement of GESA mainly depends on gradient enhancement where the AUC score is 0.936 when only utilizing gradient enhancement (model 11) on baseline. Spatial attention helps GESA to merge gradient patterns through an attention mechanism, both the gradient enhancement and attention sparsity are considered in our GESA. If just involving the spatial attention without the gradient enhancement, the performance (model 9) will encounter a significant score drop. When comparing with other frequency-based approach (model 10), our GESA utilizes smaller parameters with only a single branch architecture and achieves better performance. To better understand the effectiveness of GESA, the visualization of feature maps before and after applying gradient enhancement are shown in Fig. 5(c). Benefited from the GESA, both the facial feature and the facial contour are enhanced, which are important in forgery detection and robust prediction.

Table 6: Ablation study on FF++ c40 (low quality) to evaluate the effects of components.

ID	Methods	AUC	$PUP_{0.5}$
1	Xception (Baseline)	0.917	65.9%
2	+ Simple Stack	0.923	48.0%
3	+ 3D Conv [51]	0.931	43.6%
4	+ Channel & Spatial Attention [22]	0.929	46.9%
5	+ VRCA & GESA (ours)	0.951	34.4%
6	+ Channel Attention [22]	0.927	49.4%
7	+ Channel Variance	0.933	39.7%
8	+ VRCA	0.939	35.6%
9	+ Spatial Attention [46]	0.926	52.0%
10	+ Frequency [41]	0.933	-
11	+ Gradient Enhancement	0.936	41.0%
12	+ GESA	0.940	37.3%

5 CONCLUSION

In this paper, we propose an innovative attention-based framework, names as RobustForensics to eliminate the prediction instability in forgery detection. The proposed RobustForensics consists of two modules, *Variance-Restricted Channel Activation (VRCA)* and *Gradient-Enhanced Spatial Attention (GESA)*. The VRCA correlates channel-wise features among input frames with an adaptive activation based on variance. The GESA enhances the edge information and increases the sparsity of attention map through the spatial neighbourhood gradient. Extensive experiments demonstrate the effectiveness and significance of our approaches in in-domain detection, generalization on distortions and generalization on zero-shot learning. Although our method has outstanding effect on forgery video, for perturbed fake images or videos, our method usually fails to make stable predictions, which is caused by black-box attacks on the model. In the follow-up work, we hope to further improve the performance of the model, so that the forged video of black-box attack can also be stably detected.

REFERENCES

[1] [n.d.]. Deepfakes. https://github.com/deepfakes/faceswap/.

[2] [n.d.]. Faceswap. https://github.com/MarekKowalski/FaceSwap/.

[3] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–7.

[4] Belhassen Bayar and Matthew C Stamm. 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*. 5–10.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.

[6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).

[7] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[8] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. 2020. What makes fake images detectable? Understanding properties that generalize. *arXiv preprint arXiv:2008.10588* (2020).

[9] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2020. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364* (2020).

[10] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. 2019. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8351–8361.

[11] Zehao Chen and Hua Yang. 2020. Manipulated Face Detector: Joint Spatial and Frequency Domain Attention Network. *arXiv preprint arXiv:2005.02958* (2020).

[12] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. 2020. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10663–10671.

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

[14] Davide Cozzolino, Diego Gragnaniello, and Luisa Verdoliva. 2014. Image forgery localization through the fusion of camera-based, feature-based and pixel-based techniques. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 5302–5306.

[15] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 2017. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*. 159–164.

[16] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5781–5790.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[18] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. 2019. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686* (2019).

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*. 6202–6211.

[20] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. 2012. Image forgery localization via fine-grained analysis of CFA artifacts. *IEEE Transactions on Information Forensics and Security* 7, 5 (2012), 1566–1577.

[21] Jessica Fridrich and Jan Kodovsky. 2012. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security* 7, 3 (2012), 868–882.

[22] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3146–3154.

[23] Teddy Surya Gunawan, Siti Amalina Mohammad Hanafiah, Mira Kartiwi, Nanang Ismail, Nor Farahidah Za'bah, and Anis Nurashikin Nordin. 2017. Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)* 7, 1 (2017), 131–137.

[24] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[27] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[28] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2886–2895.

[29] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. 2018. Detection of deep network generated images using disparities in color components. *arXiv preprint arXiv:1808.07276* (2018).

[30] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5001–5010.

[31] Yuezun Li and Siwei Lyu. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656* (2018).

[32] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3207–3216.

[33] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. 2020. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8060–8069.

[34] Daniel Mas Montserrat, Hanxiang Hao, Sri K Yarlagadda, Sriram Baireddy, Ruiting Shao, Janos Horvath, Emily Bartusiak, Justin Yang, David Guera, Fengqing Zhu, et al. 2020. Deepfakes Detection with Automatic Face Weighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 668–669.

[35] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. 2020. Two-branch Recurrent Network for Isolating Deepfakes in Videos. *arXiv preprint arXiv:2008.03412* (2020).

[36] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 83–92.

[37] Scott McCloskey and Michael Albright. 2018. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247* (2018).

[38] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876* (2019).

[39] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467* (2019).

[40] Xunyu Pan, Xing Zhang, and Siwei Lyu. 2012. Exposing image splicing with inconsistent local noise variances. In *2012 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–10.

[41] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues. *arXiv preprint arXiv:2007.09355* (2020).

[42] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2017. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.

[43] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1–11.

[44] Irwin Sobel. 2014. An Isotropic 3x3 Image Gradient Operator. *Presentation at Stanford A.I. Project 1968* (02 2014).

[45] Luchuan Song, Bin Liu, Guojun Yin, Xiaoyi Dong, Yufei Zhang, and Jia-Xuan Bai. 2021. TACR-Net: Editing on Deep Video and Voice Portraits. In *Proceedings of the 29th ACM International Conference on Multimedia*. 478–486.

[46] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. 2021. Bottleneck Transformers for Visual Recognition. *arXiv preprint arXiv:2101.11605* (2021).

[47] Jingqun Tang, Wenqing Zhang, Hongye Liu, MingKun Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. 2022. Few Could Be Better Than All: Feature Sampling and Grouping for Scene Text Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4563–4572.

[48] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

[49] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.

[50] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. 2020. DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance. *arXiv preprint arXiv:2004.07532* (2020).

[51] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[53] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. 2020. High-frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8684–8694.

[54] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[55] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 7.

[56] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

[57] Yaohui Wang and Antitza Dantcheva. 2020. A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes. In *FG'20, 15th IEEE International Conference on Automatic Face and Gesture Recognition, May 18-22, 2020, Buenos Aires, Argentina.*

[58] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8261–8265.

[59] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. 2020. Exploring Self-attention for Image Recognition. In *CVPR*.

[60] Minghang Zheng, Peng Gao, Xiaogang Wang, Hongsheng Li, and Hao Dong. 2020. End-to-End Object Detection with Adaptive Clustering Transformer. *arXiv preprint arXiv:2011.09315* (2020).

[61] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2020. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *arXiv preprint arXiv:2012.07436* (2020).

[62] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. 2017. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1831–1839.