Learning Pairwise Interaction for Generalizable DeepFake Detection

Ying Xu[1], Kiran Raja[1], Luisa Verdoliva[2], Marius Pedersen[1]
[1] Norwegian University of Science and Technology, Norway
{ying.xu, kiran.raja, marius.pedersen} @ntnu.no
[2] University Federico II of Naples, Italy verdoliv@unina.it

Abstract

A fast-paced development of DeepFake generation techniques challenge the detection schemes designed for known type DeepFakes. A reliable Deepfake detection approach must be agnostic to generation types, which can present diverse quality and appearance. Limited generalizability across different generation schemes will restrict the wide-scale deployment of detectors if they fail to handle unseen attacks in an open set scenario. We propose a new approach, Multi-Channel Xception Attention Pairwise Interaction (MCX-API), that exploits the power of pairwise learning and complementary information from different color space representations in a fine-grained manner. We first validate our idea on a publicly available dataset in a intra-class setting (closed set) with four different Deepfake schemes. Further, we report all the results using balanced-open-set-classification (BOSC) accuracy in an inter-class setting (open-set) using three public datasets. Our experiments indicate that our proposed method can generalize better than the state-of-the-art Deepfakes detectors. We obtain 98.48% BOSC accuracy on the FF++ dataset and 90.87% BOSC accuracy on the CelebDF dataset suggesting a promising direction for generalization of DeepFake detection. We further utilize t-SNE and attention maps to interpret and visualize the decision-making process of our proposed network.

1. Introduction

Deepfakes are synthetic media that are generated by deep learning methods to manipulate the content in images and videos. The manipulations include altering people's identities, faces, expressions, speech or bodies to both entertainment and malicious intent (for example pornographic uses). Benefiting from the remarkable advancement in generation models, amateurish individuals are capable of creating Deepfakes using off-the-shelf models [1, 4, 3] without tedious efforts. In the meantime, channelized efforts have been dedicated to devising Deepfakes detection algorithms using multiple approaches such as by determining unique artifacts [37, 15, 23, 25, 8, 32], utilizing Convolutional Neural Networks (CNNs) based networks [36, 41, 39], employing frequency domain information [22, 13, 40] and other clues [18, 16].

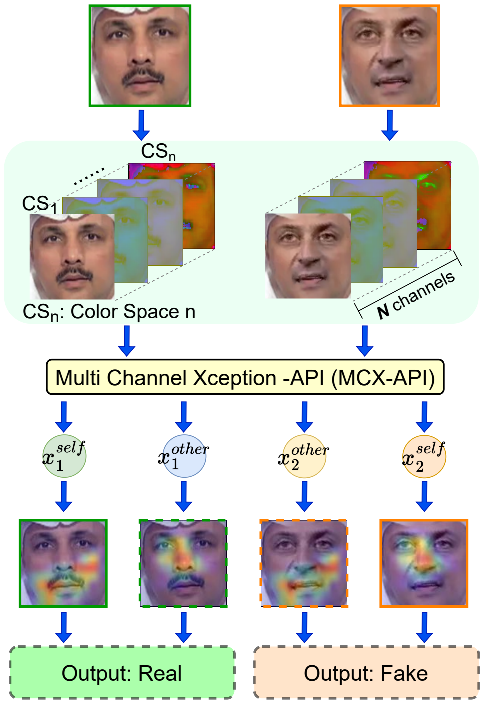


Figure 1: Overview of proposed Multi Channel Xception Attentive Pairwise Interaction (MCX-API) network. Two inputs are first represented in n kinds of color spaces, CS_1 to CS_n to obtain a two N-channel input and subsequently feature vectors. We thereafter obtain x_i^{self} and x_i^{other} by comparison through MCX-API, where x_i^{self} is enhanced by its own images and x_i^{other} is activated by the other image. x_i is therefore improved with discriminative clues that come from both images. By comparison, we can finally distinguish if an image is pristine or fake.

With an atomic effort, these methods could perform well with an average of more than 99% [41] accuracy in a closed-set problem where the training and testing data are pulled from the same label and feature spaces. For example, the network is trained on attacks A, B and C and tested on images/videos drawn from attack A or B or C. However, newer DeepFakes generation mechanisms make the detection algorithms untrustworthy and non-generalizable by degrading the performance of the detector [55, 57] as no exception to those classifiers trained with machine learning methods. In the context of DeepFakes detection, this can be parallel to detecting attack D when the detector is trained on A, B, and C, making it an open-set problem. The reasons behind the collapse of detection models towards unseen contents can, to some degree, be attributed to various generation algorithms, which often result in different data distributions, feature spaces, and appearance properties of images or videos. While one can see the imperative need for a generalizable detection technique to make reliable decisions on unknown/unseen generation types in addition to known/seen generation data, we note low performances of networks in this direction [55, 57, 51, 10].

We thus motivate our work, focusing on both closed-set and open-set detection in this article. We draw our inspiration from how humans tend to detect altered media in a fine-grained manner by comparing one kind of visual content to another. Human decision making relies on detecting an unseen kind of manipulated images/videos as fake by comparing the unknown generation type to the known generation types, especially the artifacts and clues [58]. Initial work using on pairwise interaction has shown promising directions to capture subtle differences in a pairwise manner with not only principal parts of the image but also distinct details from the other image [58]. Using such a paradigm, we propose to learn the known type of generations in a fine-grained pairwise manner explicitly to improve the performance of a Deepfake detector for unknown types. Further, we also note the complementary information an image/video can exhibit in different color spaces along the same lines. We therefore incorporate information from four color spaces, including RGB, CIELab, HSV, and YCbCr integrating to boost the attentive pairwise learning to guide the detector to classify the non-manipulated images efficiently. Our proposed approach exploits the information from color channels in a pairwise manner using the strengths of the Xception network and we refer to this as the Multi-Channel Xception Attentive Pairwise Interaction (MCX-API) network between non-manipulated images against a set of manipulated images and to try to generalize the detector towards unknown manipulation types or unseen data. Figure 1 shows an overview of the idea presented in this work.

To validate our idea in this work, we conduct various experiments using FaceForensics++ dataset [41] which consists of four different manipulation classes including DeepFakes (DF) [2], FaceSwap (FS) [4], Face2Face (F2F) [47] and NeuralTextures (NT)[46] where we obtain better state-of-the-art (SOTA) performance or at par detection performance to best performing SOTA approaches in closed-set experiments [14, 7, 55, 32]. Furthermore, we demonstrate the effectiveness of variants of the proposed approach in detecting Deepfakes in open-set scenarios where our approach achieves better results than SOTA models on three other public datasets such as FakeAV [28], KoDF [30], and Celeb-DF [35].

A detailed ablation study is presented on MCX-API to illustrate the variability of performance of the detector to various design choices in the network. Thus, the main contributions of our paper are **(1)** We propose a new framework - Multi-Channel Xception Attentive Pairwise Interaction (MCX-API) for Deepfakes detection by exploiting color space and pairwise interaction simultaneously, bringing a novel fine-grained idea for the Deepfakes detection field. **(2)**We report all results by balanced-open-set-classification (BOSC) accuracy to exemplify the generalizability of our proposed approach. **(3)**We conduct cross-datasets validations with three SOTA Deepfake datasets, Celeb-DF [35], KoDF [30] and FakeAVCelebDF [28]. Furthermore, we compared the results with SOTA Deepfake detection methods. Our MCX-API obtains 98.48% BOSC accuracy on the FF++ dataset and 90.87% BOSC accuracy on the Celeb-DF dataset, indicating an optimistic direction for the generalization of DeepFake detection.

In the rest of the paper, we list a set of directly related works in Sec. 2 and then present our proposed approach in Sec. 3. We provide an analysis of explainability in Sec. 5 with the set of experiments and results on generalizability detailed in Sec. 4. We finally conclude the work in Sec. 7.

2. Related Work

Deepfakes detection methods. A track of Deepfakes detection focuses on the unique artifacts on human faces, such as eye blinking [33], different eye colors [37], abnormal heartbeat rhythms shown on the face [15, 23]. LipForensics [25] targets high-level semantic abnormalities in mouth movements, which the authors observe as a common indicator in many generated videos. Some articles are dedicated to finding inconsistencies in images and videos. These inconsistencies arise out of generation process where landmarks, head pose are inconsistent [52, 8] or observable in image blending [34, 32]. Cozzolino *et. al.* [18] have introduced ID-Reveal, an identity-aware detection approach leveraging a set of reference videos of a target person and trained in an adversarial manner. Many papers have utilized CNNs-based methods for detecting features existing in forged images[36, 41, 39]. Using high-frequency features [22, 13, 40] to distinguish Deepfakes are also gaining

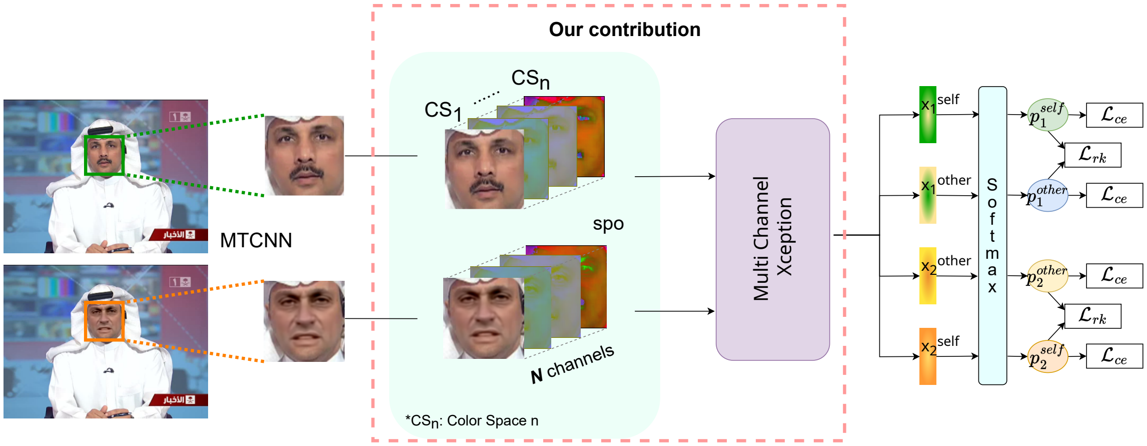


Figure 2: The architecture of MCX-API network.

more popularity on this topic. Although pairwise learning have been used for Deepfake detection [26, 51], they lack the pairwise interactions by using contrastive learning.

Generalization to unseen manipulations. While many works are proposed for detecting DeepFakes, they have focused on closed-set experiments where the training and testing set distributions are similar. The open-set experiments indicate that they underperform on unseen manipulations. In the meantime, an increasing number of works have tried to address the problem of generalization of DeepFakes detection. These works have focused on domain adaptation and transfer learning to minimize the task of learning parameters in an end-to-end manner [10, 29, 31]. *Cozzolino et. al.* [19] proposed an autoencoder-like structure ForensicTransfer and the generalization aspect was studied using a single detection method for multiple target domains. The follow-up works like Locality-aware AutoEncoder (LAE) [21] and Multi-task Learning were proposed for detecting and segmenting manipulated facial images and videos [38]. Transfer learning-based Autoencoder with Residuals (TAR) [31] recently proposed uses the residuals from autoencoders to handle generalizability. *Kim et. al.* [29] employed the Representation Learning (ReL) and Knowledge Distillation (KD) paradigms to introduce a transfer learning-based Feature Representation Transfer Adaptation Learning (FReTAL) method. While these transfer learning and zero-shot/few-shot learning methods could not wholly deal with the Deepfakes detection generalization problem, because the networks have already seen the manipulated image/videos. Therefore, strictly speaking, it is not an open-set situation.

In the meantime, some other novel networks have been proposed dealing with the generalization problem of Deepfakes detection. A new method to detect deepfake images using the cue of the source feature inconsistency within the forged images [55] is proposed based on the hypothesis that distinct source features can be preserved and extracted through SOTA deepfake generation processes. Joint Audio-Visual Deepfake Detection [57] is proposed by jointly modeling video and audio modalities. This novel visual/auditory deepfake combined detection task shows that exploiting the intrinsic synchronization between the visual and auditory modalities could benefit deepfake detection. *Xu et. al.* [51] proposed a novel method using supervised contrastive learning to deal with the generalization problem in detecting forged visual media.

3. Proposed Method

Fine-grained method has been widely used for classification problems where the categories are visually very similar [58, 50, 9]. We draw similar inspiration to our problem of Deepfake detection following the architecture proposed by earlier [58] and build upon with number of improvements. We assert that architecture for fine-grained classification can help in detecting Deepfakes. Unlike the orginal architecture, we introduce Xception [14] to extract the embeddings motivated by earlier works in Deepfake detection [41, 54, 49, 29].

Second, to benefit from information from different color spaces, we make the base network to a multi-channel network. Then, we enforce pairwise learning by following the architecture of Attentive Pairwise Learning [58]. We propose using the Multi Channel Xception Attentive Pairwise Interaction Network (MCX-API) to deal with the Deepfakes classification problem as detailed further.

3.1. Architecture

We first utilize MTCNN[53] to crop and align the face region of a single frame. Two selected face images are further sent to a Multi-Channel Xception backbone, and this backbone network extracts two corresponding D-dimension feature vectors x_1 and x_2 using the face image represented in N different channels that include RGB, CIELab, HSV, and YCbCr. A mutual vector $x_m \in \mathbb{R}^D$ is further generated by concatenating x_1 and x_2 and using a Multi-Layer Perceptron (MLP) function for mapping x_m to get a D dimension. x_m is a joint feature that includes high-level contrastive clues of both input images across multiple color channels.

In order to compare x_m with x_1 and x_2, we need to activate x_m using sigmoid function to increase the positive relation with x_i and decrease the negative relation against x_i [58]. Therefore, two gate vectors g_1 and g_2 will be generated. g_i is calculated by x_m and x_i, thus containing contrastive clues and acting as discriminative attention spots semantic contrasts with a distinct view of each x_i. The gate vector g_i is the sigmoid of the output of channel-wise product between x_m and x_i, whose formula is provided in Equation (1).

$$g_i = sigmoid(x_m \odot x_i), \quad i \in \{1, 2\} \tag{1}$$

A pairwise interaction between input features x_i and gate vectors g_i is performed to induce residual attention by comparing one image to the other to distinguish the final class. The sequence of interaction can be shown in Equation (2).

$$\begin{aligned} x_1^{pristine} &= x_1 + x_1 \odot g_1 \\ x_1^{fake} &= x_1 + x_1 \odot g_2 \\ x_2^{pristine} &= x_2 + x_2 \odot g_2 \\ x_2^{fake} &= x_2 + x_2 \odot g_1 \end{aligned} \tag{2}$$

Through the pairwise interaction of each feature x_i, two attentive feature vectors $x_i^{pristine} \in \mathbb{R}^D$ and $x_i^{fake} \in \mathbb{R}^D$ are further produced. The former one is highlighted by its gate vector, and the latter is triggered by the gate vector of the compared image. x_i is thus enhanced with discriminative clues from both input features through pairwise interaction.

3.2. Loss calculation

The four attentive features x_i^j where $i \in \{1, 2\}$ and $j \in \{pristine, fake\}$, the pairwise interaction outputs, are fed into a $softmax$ classifier for the loss calculation [58]. The output of $softmax$ denoted by p_i^j is the probability of a feature belonging to a specific class (i.e., non-manipulated or Deepfake). The main loss in our case is the cross-entropy loss

$$\mathcal{L}_{ce} = -\sum_{i \in \{1,2\}} \sum_{j \in \{pristine, fake\}} y_i^\intercal log(p_i^j) \tag{3}$$

where y_i is the one-hot label for image i in the pair and $\intercal$ represents the transpose. MCX-API can be trained to determine all the attentive features x_i^j under the supervision of the label y_i through this loss.

Furthermore, a hinge loss of score ranking regularization

$$\mathcal{L}_{rk} = \sum_{i \in \{1,2\}} max(0, p_i^{fake}(c_i) - p_i^{pristine}(c_i) + \epsilon) \tag{4}$$

is also introduced when computing the complete loss [58]. c_i is the corresponding index associated with the ground truth label of image i. So $p_i^j(c_i)$ is a softmax score of p_i^j. Since $x_i^{pristine}$ is activated by its gate vector g_i, it should contain more discriminative features to identify the corresponding image, compared to x_i^{fake}. $\mathcal{L}_{rk}$ is utilized to promote the priority of $x_i^{pristine}$ where the score difference between $p_i^{fake}(c_i)$ and $p_i^{pristine}(c_i)$ should be greater than a margin. The whole loss for a pair is composed of two losses, cross-entropy loss $\mathcal{L}_{ce}$ and score ranking regularization $\mathcal{L}_{rk}$ with coefficient λ.

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{rk} \tag{5}$$

In this way, MCX-API is able to take feature priorities into account adaptively and learns to recognize each image in the pair.

4. Experiments and Results

4.1. Datasets

Training data: We select FaceForensics++ [41] to train the proposed approach. This forensics dataset consists of 1000 original videos and corresponding number of manipulated videos consisting of 1000 videos for each of the subsets - DeepFakes (denoted as DF) [2], Face2Face (denoted as F2F) [47], FaceSwap (denoted as FS) [4], and NeuralTextures (denoted as NT) [46].

Cross-dataset Validation: We also select three other SOTA datasets for generalization test and comparison. **Celeb-DF [35]**: For Celeb-DF, we choose id51-id61 from Celeb-real, Celeb-synthesis and id240-id299 from YouTube-real for the test set. **KoDF [30]** We randomly selected 265 real videos and 734 fake ones as our test set. **FakeAV [28]** We randomly selected 500 videos as our test set.

Implementation details. We choose uncompressed videos for our experiments in this work using the Pytorch framework [5] to develop the models and the experiments are conducted on Python 3.6 environment on NVIDIA Tesla V100 32Gb in IDUN system owned by NTNU [43].

Multi-task Cascade Convolutional Neural Networks (MTCNN) [53] is employed for face detection and face alignment since our experiments are focused on detecting the manipulated face region alone. We allow loose cropping of the face region to capture the entire silhouette against tight cropping. The first 30 frames from each video are extracted, resulting in 150000 total images. We use random

Table 1: **Frame-level BOSC Accuracy and AUC for our proposed MCX-API networks and SOTA methods on seen data.** We compare the results with the SOTA methods on DF/F2F/FS/NT respectively. All networks are trained on the whole FF++ c23 dataset. The data of the first three methods are adopted from Table 5 in Appendix of FF++ [14].

FF++ c23	Frame-level (BOSC(%)/AUC)				
Method	DF	F2F	FS	NT	Average
Cozzolino *et al.* [17]	75.51/ -	86.34/ -	76.81/ -	75.34/ -	78.50/ -
Bayar and Stamm [11]	90.25/ -	93.96/ -	87.74/ -	83.69/ -	88.91/ -
MesoNet [7]	89.55/ -	88.60/ -	81.24/ -	92.19/ -	87.90/ -
Xception*[14]	96.35/0.9941	96.26/0.9937	96.29/0.9952	92.43/0.9736	95.33/0.9892
SupCon*[51]	97.18/0.9984	96.88/0.9978	97.05/0.9980	92.92/0.9846	96.01/0.9947
API-Net(ResNet101)*[58]	88.71/0.9820	90.13/0.9860	87.79/0.9728	82.96/0.9248	87.40/0.9664
Ours					
MCX-API(RGB)	**98.75**/0.9996	**99.90**/0.9986	**98.5/0.9993**	**96.75**/0.9896	**98.48**/0.9968
MCX-API(RGB+HSV)	**98.75**/0.9988	98.50/0.9979	97.75/0.9978	95.75/0.9829	97.69/0.9943
MCX-API(RGB+CIELab)	97.00/0.9996	96.50/0.9985	96.25/0.9989	95.25/0.9909	96.25/0.9970
MCX-API(RGB+YCbCr)	98.00/**0.9998**	98.25/**0.9991**	97.75/**0.9993**	**96.75**/0.9920	97.69/**0.9976**
MCX-API(RGB+HSV+CIELab)	96.50/0.9990	95.50/0.9888	96.00/0.9835	95.50/**0.9933**	95.88/0.9912
MCX-API(RGB+LAB+YCbCr)	92.00/0.9963	92.25/0.9972	91.50/0.9960	91.00/0.9870	91.69/0.9941

[*] Our implementation of the method.

cropping in the training phase and center cropping during the testing phase ($512^2 \rightarrow 448^2$). In all our experiments, we employ Xception as the backbone where we derive the feature vector $x_i \in \mathbb{R}^{2048}$ after the global average pooling. We use a batch sampler during the training by randomly sampling three categories in each batch. For each category, we randomly choose nine images due to the limitations of the GPU and memory constraints. We further exercise care to have no sample overlap among all batches, as we exclude the selected sample from the dataset. We locate its most similar image from both its own class and the rest classes for each image by calculating the distance between features by utilizing both Euclidean distance and cosine distance. Each image would get one image as its intra- and inter-pair in the batch, respectively. Each pair is used as input x_1 and x_2 as well as generating a mutual vector $x_m \in \mathbb{R}^{2048}$ through the concatenation and the multilayer perceptron (MLP).

Based on empirical evaluations, we adopt the coefficient λ in Equation (5) as 1.0, and 0.05 as the margin value in the score-ranking regularization. We use cosine annealing strategy to alter the learning rate starting from 0.01 [55]. We train the network with 100 epochs and freeze the parameters in the CNN backbone, and further on train only the classifier in the first eight epochs.

Evaluation Metrics. We adopt Balanced-Open-Set-Classification (BOSC) accuracy and AUC as evaluation metrics. $BOSC = \frac{Sensitivity+Specificity}{2}$, where $Sensitivity = \frac{TP}{TP+FN}$ and $Specificity = \frac{TN}{TN+FP}$.

Table 2: Comparison of the test results on the FF++ dataset with c23 (high-quality compression) settings. Training for all networks is carried out on FF++ c23. The accuracy and AUC score are at frame-level. The best performances are marked in bold. Data for Xception, F^3-Net, and EfficientNet-B4 are adopted from Table 2 in MaDD [54].

Method	ACC	AUC
Xception	95.73	0.9909
F^3-Net [40]	97.52	0.9810
EfficientNet-B4 [45]	96.63	0.9918
DCL [44]	96.74	0.9930
MaDD [54]	97.60	0.9929
M2TR [49]	97.93	0.9951
API-Net	87.40	0.9664
Ours	**98.48**	**0.9968**

4.2. Experimental Results

We evaluate the effectiveness of the proposed MCX-API network with both seen and unseen data in this section.

4.2.1 Intra-dataset Evaluation (Closed Set Protocol)

We conduct experiments on six networks with different color spaces on MCX-API whose results are presented in Tab. 1. All networks are trained with all four manipulated methods along with pristine in FF++ c23 dataset. We test the frame-level detection performance on the test data of FF++ c23 in a non-overlapping manner regarding the ID.

In Tab. 1, the frame-level test results are listed. We ob-

Table 3: **Video-level BOSC Accuracy and AUC for our proposed MCX-API networks and SOTA methods on unseen data.** We compare the results with the SOTA methods on FakeAV/KoDF/Celeb-DF respectively. All the networks are trained on the whole FF++ c23 dataset. The data of the SOTA methods are adopted from Table 2 from [16].

FF++ c23	Video-level (BOSC(%)/AUC)		
Method	FakeAV	KoDF	Celeb-DF
Xception*	23.99/0.450	25.97/0.482	31.34/0.505
Seferbekov [6]	**95.0/0.986**	79.2/0.884	75.3/0.860
FTCN [56]	64.9/0.840	63.0/0.765	-
LipForensics [25]	83.3/0.976	56.1/0.929	-/0.820
ID-reveal [18]	63.7/0.876	60.3/0.702	71.6/0.840
POI [16]	86.6/0.941	81.1/0.899	-
API-Net(ResNet101)*	59.99/0.72	66.92/0.76	58.00/0.76
Ours			
MCX-API(RGB)	74.94/0.95	78.09/0.87	77.88/0.87
MCX-API(HSV)	74.63/0.75	80.64/0.85	75.67/0.88
MCX-API(CIELab)	84.28/0.90	**81.16/0.90**	64.28/0.81
MCX-API(RGB+HSV)	71.58/0.93	78.11/0.87	80.18/0.88
MCX-API(RGB+CIELab)	83.89/0.93	77.93/0.83	68.34/**0.91**
MCX-API(RGB+YCbCr)	70.41/0.92	78.39/0.85	**90.87**/0.90
MCX-API(RGB+HSV+CIELab)	92.38/0.98	78.91/0.83	59.04/0.89
MCX-API(RGB+LAB+YCbCr)	82.93/0.96	76.20/0.80	54.92/0.85

* Our implementation of the method.

serve that our proposed MCX-API network with RGB inputs reaches the highest average accuracy, 98.48%. In addition, this setting also gains the highest accuracy on DF, F2F, and FS with 98.87%, 99.90% and 98.50%, respectively. MCX-API with YCbCr achieves the highest accuracy for NT with 97.00%. As RGB provides best performance under 3-channel setting, we combine RGB with HSV, CIELab, and YCbCr, respectively, to create three 6-channel MCX-API networks. From the second block in Tab. 1, we can see that RGB+YCbCr obtains the highest average AUC score of 0.9976 and the best performance on DF, F2F, and FS regarding AUC score. This indicates better prediction output scores using MCX-API with the combination of RBG and YCbCr color spaces. The 9-channel MCX-API network with RGB, HSV, and CIELab further gains the highest 0.9933 AUC score for NT.

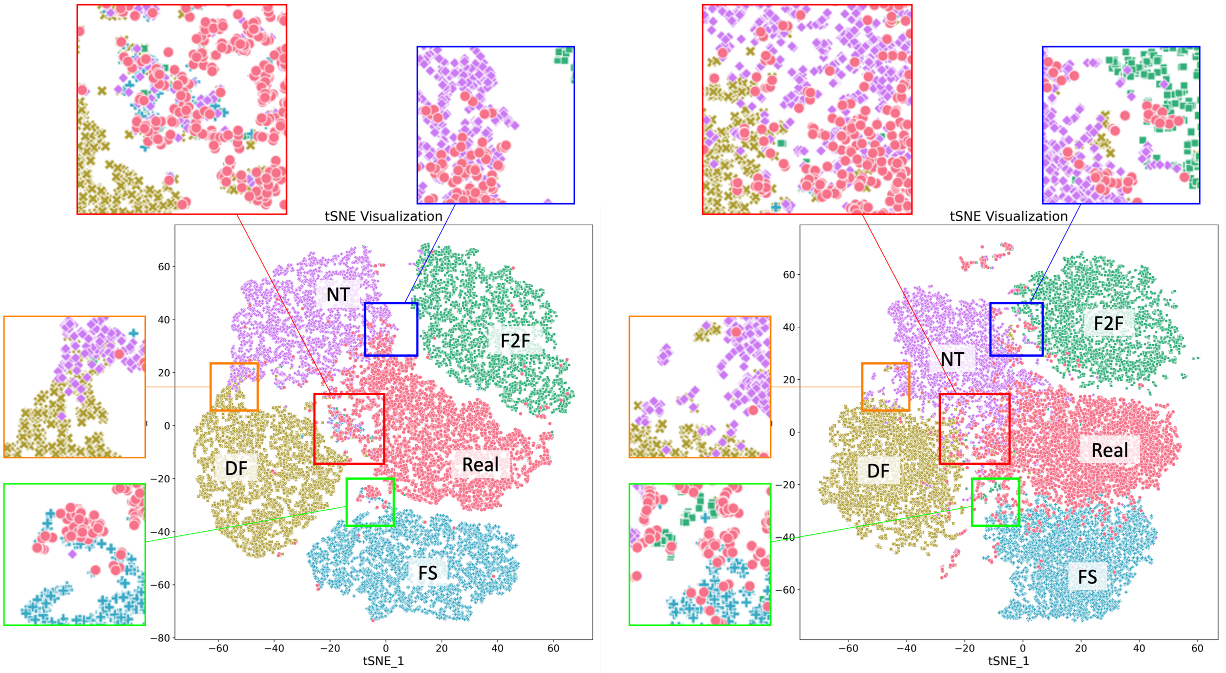
The results of the comparison with the SOTA methods are reported in Tab. 2. All networks are trained on FF++ c23 (high-quality compression). The accuracy and AUC scores are measured at frame level. The results are averaged on all the test sets from FF++ c23, including pristine and all four kinds of manipulated videos. Our proposed method MCX-API with RGB color space obtains the best performance compared to SOTA methods. The best accuracy of the BOSC is 98.48%, and the highest AUC score is 0.9968. The result shows that our idea of pairwise learning in a fine-grained manner could work well in inter-class (closed-set) setting of Deepfake detection problem.

4.2.2 Cross-dataset Evaluation

We conduct a comparison on cross-dataset validation with SOTA methods to validate the proposed approach. We employ FakeAV, KoDF, and Celeb-DF to test the generalizability of our MCX-API network. Training for all networks are carried out on the FF++ c23 dataset and tested on FakeAV, KoDF, and Celeb-DF. We note that MCX-API with CIELab color space gets the best scores for KoDF with an accuracy of 81.86% and an AUC score of 0.90 as presented in Tab. 3. MCX-API with RGB+YCbCr wins in the cross-dataset validation for Celeb-DF with an accuracy of 90.87% and the second best AUC score 0.90. MCX-API with color space RGB+HSV+CIELab achieves the second best place for FakeAV with 92.38% accuracy and 0.98 AUC score. In general, our proposed network gets a relatively better performance than the SOTA methods which indicates the better generalizability of the proposed MCX-API network.

5. Explainable Analysis of MCX-API

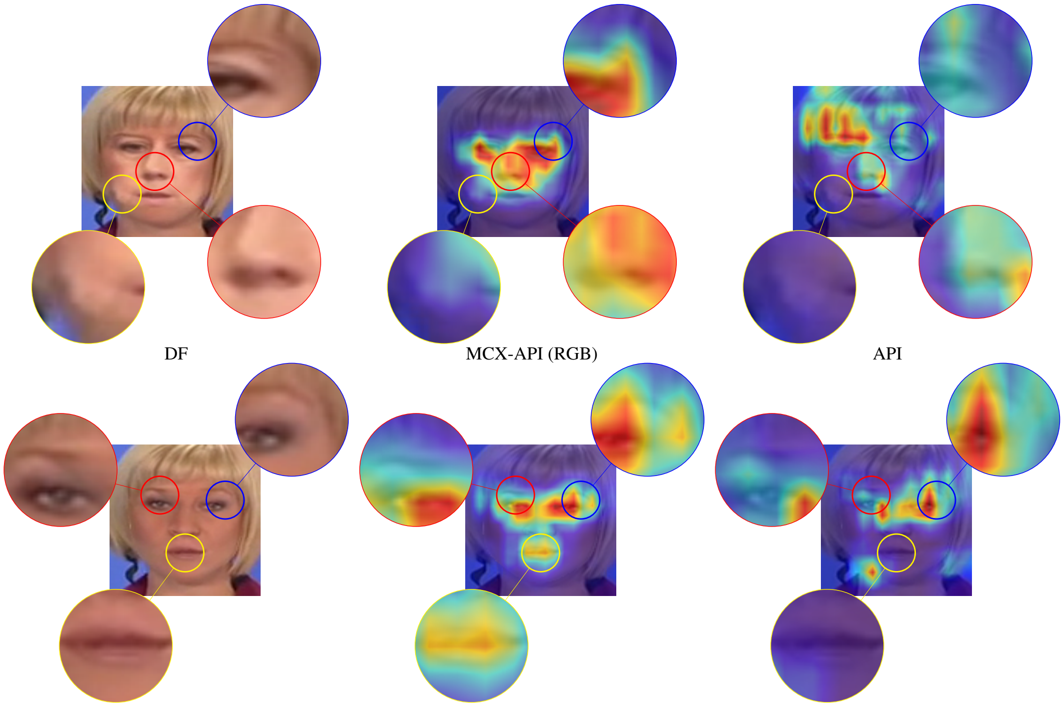
We further analyze the network to understand the performance gain by analyzing embeddings using t-SNE plots [48] and class activation maps [42, 12, 20, 27, 24]. While the t-SNE provides topology explanations of the learned features, the activation maps allow for a better visualization of what has been learned by our network.



MCX-API (RGB)

API

Figure 3: Data visualizations in 2D by t-SNE for MCX-API(RGB) and API. The left plot is t-SNE for our proposed MCX-API. The right plot is t-SNE for base architecture API-Net. We blow up the intersection parts and outliers for a clear view.



DF

MCX-API (RGB)

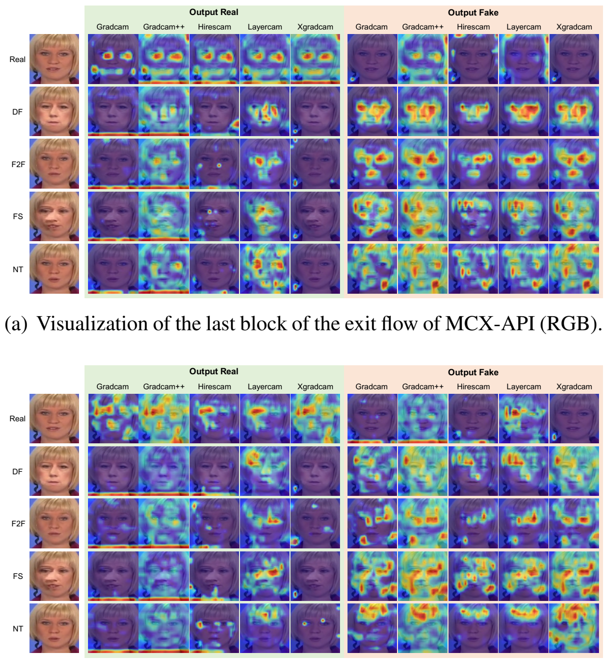
API

F2F

MCX-API (RGB)

API

Figure 4: Blow up in activation maps from LayerCAM analysis of MCX-API(RGB) and base architecture API-Net on DF and F2F faces.



(a) Visualization of the last block of the exit flow of MCX-API (RGB).

(b) Visualization of the last block of the API-Net.

Figure 5: Visualization of the last layer of MCX-API (RGB) and API networks. We utilize Grad-CAM [42], Grad-CAM++ [12], HiResCAM [20], LayberCAM [27] and XGradCAM [24] as our visualization tool. For larger figure, please refer to Fig. 6.

5.1. Data Visualizations With t-SNE

The results of a t-SNE 2D map for the feature vectors are illustrated in Fig. 3. We compare the t-SNE of our MCX-API and base architecture API-Net. We notice that the five classes of Real, DF, F2F, FS, and NT for MCX-API are well separated with five different clusters as against the base architecture of API-Net. There is an unclear boundary between Real and NT, shown in the blue box for MCX-API. This overlapping can be the reason for the relatively lower accuracy obtained on NT. There are small areas overlapping between DF/NT(yellow/purple) and Real/F2F(red/blue). We further notice a few samples of Real (red dots) distributed in each fake class, leading to the errors of our proposed network.

5.2. Visualizing Decisions With Attention Maps

We apply different class activation visualization methods on the last layer of proposed network to analyze MCX-API shown in Fig. 5. For comparison, we also show the visualization of the base API-Net. Precisely, we adopt Grad-CAM [42], Grad-CAM++ [12], HiResCAM [20], LayberCAM [27] and XGradCAM [24]. The visualization results are provided in Fig. 6(a) for our proposed MCX-API and in Fig. 6(b) for API-Net.

The activation map for Output Real is on the left part with a green background, and the activation map for Output Fake is on the right part with a pink background. The rows from top to bottom are the visualization for five classes of Real, DF, F2F, FS, and NT, respectively. We can observe that real images gains more attention within Output Real(left part) than Output Fake(right part). In contrast, fake images obtain more attention within Output Fake than Output Real. This explains the ability of our network to detect Deepfakes.

We further blow up the activation maps from LayerCAM for DF and F2F images in Fig. 4. From visual analysis, it is evident that the MCX-API focuses more on the facial region, such as the eyes and the mouth. For instance, double eyebrows are found in the DF image (blue circle). MCX-API pays more attention than API around this region.

6. Limitations of our work

We notice in Tab. 1 that with the increase in color spaces, there are no apparent improvements in BOSC accuracy. We assume that there is redundant information among channels, and further work would be focused on finding the most helpful color information to extend our proposed approach. We also observe that no single configuration could perform reasonably well for all the unseen data, which is the biggest issue for Deepfake detection field. Introducing other information, such as temporal data and audio, would be a good idea as more inconsistency could be found by extending our work to video based approach.

7. Conclusion

There is an imperative need for a generalized DeepFakes detection method to deal with the newer manipulation methods in visual media. In this paper, we proposed to apply the Multi-Channel Xception Attentive Pairwise Interaction (MCX-API) network to the Deepfakes detection field in a fine-grained manner. We conducted experiments on the publicly available FaceForensics++ dataset, and our approach obtained better performance than the SOTA approaches on both seen and unseen manipulation types. We obtain 98.48% BOSC accuracy on the FF++ dataset and 90.87% BOSC accuracy on the CelebDF dataset suggesting a promising direction for the generalization of DeepFake detection. Comprehensive ablation studies have been conducted to understand our algorithm better. We further explain the performance of our network by using t-SNE and attention maps. The results showed that Deepfake had been well separated from real videos. While our approach has indicated a promising solution to obtain a generalized detection mechanism, we have listed certain limitations that can pave the way for future work.

References

[1] Deepfacelab. https://github.com/iperov/DeepFaceLab.

[2] Deepfakes-faceswap. https://github.com/deepfakes/faceswap. 2021-10-25.

[3] Faceapp - most popular selfie editor. https://www.faceapp.com/.

[4] Marekkowalski-faceswap. https://github.com/MarekKowalski/FaceSwap. 2021-10-25.

[5] Pytorch. https://pytorch.org/.

[6] Seferbekov, s.: Deepfake detection (dfdc) team sefer. https://github.com/selimsef/dfdc_deepfake_challenge/.

[7] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[8] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, 2019.

[9] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2927–2936, 2015.

[10] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint arXiv:2006.11863*, 2020.

[11] Belhassen Bayar and Matthew C Stamm. 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*, pages 5–10, 2016.

[12] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[13] Zehao Chen and Hua Yang. Attentive semantic exploring for manipulated face detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1985–1989. IEEE, 2021.

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

[15] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[16] Davide Cozzolino, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. *arXiv preprint arXiv:2204.03083*, 2022.

[17] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 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*, pages 159–164, 2017.

[18] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021.

[19] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018.

[20] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv e-prints*, pages arXiv–2011, 2020.

[21] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards generalizable forgery detection with locality-aware autoencoder. 2019.

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

[23] Steven Fernandes, Sunny Raj, Eddy Ortiz, Iustina Vintila, Margaret Salter, Gordana Urosevic, and Sumit Jha. Predicting heart rate variations of deepfake videos using neural ode. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[24] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

[25] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021.

[26] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee. Deep fake image detection based on pairwise learning. *Applied Sciences*, 10(1):370, 2020.

[27] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[28] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: a novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[29] Minha Kim, Shahroz Tariq, and Simon S Woo. Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1012, 2021.

[30] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10744–10753, 2021.

[31] Sangyup Lee, Shahroz Tariq, Junyaup Kim, and Simon S Woo. Tar: Generalized forensic framework to detect deep-

fakes using weakly supervised learning. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 351–366. Springer, 2021.

[32] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[33] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

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

[35] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[36] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019.

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

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

[40] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020.

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

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[43] Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure, 2019.

[44] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2316–2324, 2022.

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

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

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

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[49] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 615–623, 2022.

[50] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 842–850, 2015.

[51] Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389, 2022.

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

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[54] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021.

[55] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021.

[56] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15044–15054, 2021.

[57] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021.

[58] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13130–13137, 2020.