Deepfake Catcher: Can a Simple Fusion be Effective and Outperform Complex DNNs?

Akshay Agarwal[1] and Nalini Ratha[2]
[1]IISER Bhopal, India and [2]University at Buffalo, USA
akagarwal@iiserb.ac.in and nratha@buffalo.edu

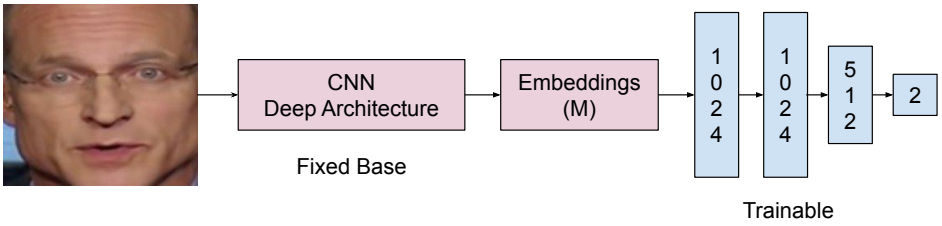
Abstract

Despite having completely different configurations, deep learning architectures learn a specific set of features that are common across architectures. For example, the initial few layers learn the low-level edge features from the images. Based on this fact, in this research, we have showcased the potential of deep neural network fusion for simple and effective deepfake detection. The advantage of building an architecture in such a manner is to build a low-power-consuming and accurate defense that can be deployed on mobile devices. To utilize the pre-trained knowledge and obtain downstream task-specific knowledge, we have identified a breakpoint in different networks and divided the obtained knowledge of a network into fixed and adaptive information. We have kept the fixed knowledge intact while modifying the adaptive knowledge along with entirely new knowledge for the deepfake detection task. In the end, the decision of multiple deep architectures trained based on their breakpoint are combined for improved performance. Extensive comparisons performed with existing state-of-the-art architectures demonstrate the effectiveness of the proposed deepfake detection algorithm. The proposed algorithm not only surpasses the existing state-of-the-art (SOTA) algorithms but also needs low computational power. We have further challenged the proposed algorithm by evaluating it by collecting real-world deepfake images.

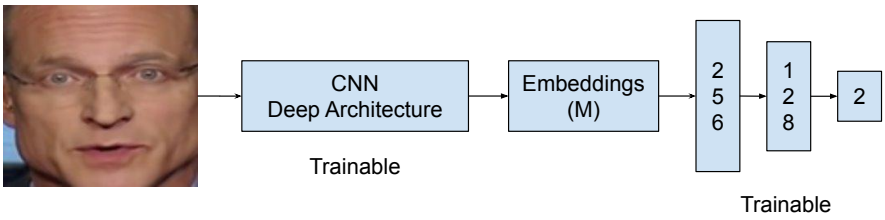
1. Introduction

Fake images especially where few components of the images are swapped can lead to a significant decrease in the image classification performance [8, 35]. Further, with the advancement in machine learning and generative networks, both the generation and modification of images including face images have become easy, and few tap-based tasks [3, 6]. These manipulated face images can not only provide illegal access but can serve as a medium to spread false and hateful information. One of the popular and powerful manipulation techniques in today's time is called Deepfake. Deepfake is a technique where deep neural network architectures are used to create synthetic media which consists of the manipulation of either expression or identity. Since the coining of the term Deepfake, several advancements have been noticed in the generation of these synthetic media which range from the generation of low-quality videos to high-quality and resolution images [15, 26, 45]. Not only the sophisticated machine learning algorithms but also the availability of several social media applications which generally operate in a tap/click fashion can also be used to generate deepfake synthetic media. The generation of deepfake from these platforms is also stealthy and can fool several machine learning and face recognition algorithms [3, 6]. *The presence of these fake media is not limited to any platform or devices and low-power devices are also huge consumer social media content. Therefore, we can assert the impact of fake videos in society due to their widespread and hence demands an effective solution.*

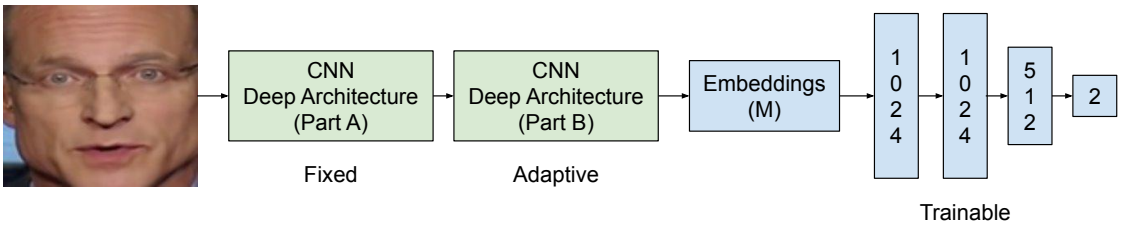
By looking at the negative impact of these synthetic media which poison social media contents, harm the security of restricted access systems, and incur personal loss including reputation and monetary aims for the development of deepfake detection algorithms. Similar to the generation of synthetic media, the development of detection algorithms has also seen tremendous growth. The detection approaches can be broadly grouped into image features equipped with traditional classifiers and fully automated deep neural networks for detection. For example, it is asserted that the physiological signals like eye blinking [30], inconsistent head poses [58], and biological signals not preserved in fake videos [15] as well as phoneme-viseme mismatches in videos [9] are the basis for detecting deepfake content. Agarwal et al. [3, 6] have proposed a novel image engineering artifacts enhancement technique for face manipulation detection. Recently due to the popularity and huge success of deep neural networks in image recognition, several deepfake face manipulation algorithms have been proposed. Motion magnified 3DNet [39], Face Warping Artifacts (FWA) [29], FWA with spatial pyramid pooling (DSP-FWA) [31], Locality



(a) Novice system (low level): fixed base knowledge and gaining new knowledge



(b) Mature system (mid level): getting new knowledge end-to-end



(c) Expert system (high level): fixed and adaptive base knowledge and gaining new knowledge

Figure 1. The concept of knowledge evolution is adopted in the proposed research. In the novice case, the knowledge obtained through pre-training for a different task is used to make a judgment for the downstream task as well. However, in the adaptive case, the pre-trained knowledge is divided into fixed knowledge which consists of generic features (part A), and adaptive knowledge which needs to be changed (part B) for the specific downstream task.

Aware Autoencoder (LAE) [18], Face X-Ray [28] and Spatial Phase Shallow Learning (SPSL) [33] are the few effective detection architectures. FWA and DSP-FWA mimic artifacts left behind by transforming and reinserting faces in the deepfake creation pipeline. using a similar concept Face X-Ray uses a self-supervised learning framework for deepfake detection. Based on the assertion of contextual information inconsistencies Zhao et al. [62] and Nirkin et al. [43] have proposed source image features and face contextual information extraction networks for deepfake detection. The deepfake detection network proposed by Zhao et al. [61] uses the multi-attention convolution network consisting of spatial attention heads and textural feature enhancement block. Not only raw images but also their transformed counterparts and combination of video and audio have also shown significant success in image manipulation and deepfake detection [4, 5, 19, 24, 35, 64]. Combination of features from different networks whether obtained through different layer types such as attention and Siamese [11] or different input processing [63] are also explored; however, not found robust against the compression artifacts. For further details of the existing works, the readers can refer to the survey papers [40, 48, 52]. **The significant limitations of the majority of the existing defense works are three folds: (i) extensive computational load both in terms of a large amount of supervised data, (ii) time to optimize million or billion of parameters, and (iii) unseen dataset generalization** [5, 55, 61, 62].

In this research, by taking these limitations seriously, we propose a deepfake detection architecture that is not only computationally efficient in terms of training but needs less labeled data and is generalized across multiple datasets and real-world videos in unseen testing settings. Fig. 1 shows the gist of the proposed deepfake detection algorithm. The proposed deep neural network architecture is inspired by the fact that different networks learn the same set of features but for a different amount of time in the network. In our case, it can be seen as a fixed knowledge of a network before its breakpoint. On top of that adaptive knowledge needs to be tweaked, and new features which are specific

to a downstream task are stacked together. Once the different breakpoints identified networks are trained, their decisions are combined to improve the overall performance. The comparisons performed with several sophisticated and computationally heavy architectures show the proposed architecture is not only towards green artificial intelligence but also achieves either state-of-the-art or comparable performances on multiple databases.

2. Proposed Algorithm

In this research, we have showcased the potential of pre-trained networks by effectively adapting them for deepfake detection and fusing their decision for state-of-the-art performance. In the literature, hundreds of deep learning architectures are proposed which consist of different layer setups ranging from sequential to residual [10, 34]. However, the majority of these networks which are pre-trained on ImageNet datasets learn the same set of features, especially in their beginning [23, 59, 60] and are independent of the class labels. Due to being generic, the modification of these features does not make sense; however, the higher-level features are specific to the class information and hence need to be adjusted for different downstream tasks.

Inspired by this fact, in this research, we have proposed the concept of breakpoints of different networks, where the layers before the breakpoint contain these generic features. The layers after the breakpoint consist of mid and high-level shapes and object-related features which are iteratively adapted for the deepfake detection task. In the end, new knowledge containing random layers is added which aims to learn the deepfake detection features only. Since the generic feature layers might be at different breakpoints, a fusion of the decision obtained from the different architectures can boost the detection performance.

The architecture of the proposed deepfake detection algorithm is shown in Fig. 2. The proposed architecture can be broadly divided into four parts: (i) fixed-based knowledge which can directly be motivated by the fact that initial layer features are generic and class-independent, (ii) adaptive knowledge: which is the knowledge that might need to be finetuned due to bias knowledge or misinformation acquired at the time of gaining the knowledge for a specific task, and (iii) new knowledge: bias-free and downstream task-specific features, and (iv) fusion: the combination of the networks consists of contrasting architectures can boost the overall performance [7, 20]. Hence, in the proposed algorithm we combined the decision of multiple architectures by assigning weight to their decision probabilities. While we have assigned equal weight without being partial to any network, the weights can easily be tuned on the validation set. In the proposed algorithm, we have studied several deep CNN architectures including VGG-16 [47], DenseNet-121 [22], and InceptionNet [49].

The steps involved in the proposed deepfake detection algorithm training can be described as follows: ❶ the individual networks are assigned with the weights adopted on the large-scale object recognition database namely ImageNet [16]; ❷ the networks are divided into two parts at different breakpoints. For instance, the VGG-16 and MobileNet networks are broken down into two pieces at layer 10, and the DenseNet-121 model is cut out into two pieces at layer 110; ❸ the broken piece corresponds to the fixed knowledge and adaptive knowledge. The motivation for keeping a few layers fixed comes from the observation that initial layers of different architectures learn somewhat similar low-level image information, i.e., edge and lines information [23, 59, 60]. Whereas, the layers followed to learn the high-level features which can be adapted for the problem and data in hand; ❹ in each network few new layers are added to being entirely new perspective and learn task-specific features only; ❺ the above three parts: (i) fixed knowledge, (ii) adaptive knowledge, and (iii) new knowledge forms a complete architecture and then trained for deepfake detection; ❻ once each architecture is trained for deepfake detection, a final decision fusion has been performed using the following decision/score fusion equation:

$$[r, f] = w_1 \odot [r_1, f_1] + \cdots + w_n \odot [r_n, f_n]$$

where, $[r1, f1]$ and $[r_n, f_n]$ are the scores computed from the deepfake architectures belonging to real (r_1 and r_n values) and deepfake class (f_1 and f_n values). $\odot$ represents the dot product. w_1 and w_n are weights for combination for decision. The final decision on an image is taken based on the max value between final scores r and f belonging to real and deepfake classes, respectively.

2.1. Implementation Details

Each architecture is trained for 50 epochs using Adam optimizer [25] with an adaptive learning rate. The initial learning rate is set to $1e^{-4}$, batch size of 32, and categorical cross-entropy loss is used for parameter optimization. The seed variable is set to 1000 for the reproducibility of the experiments. The codes are implemented in Python using Keras [14] run through TensorFlow [1] as the backend. The NVIDIA GeForce RTX 2080 GPU machine is used with the CUDA v11+. The weight values for score fusion are set to 0.5. To compute the score, the softmax activation function has been used at the final layer in each network. The ReLU activation function has been used in the intermediate layers of the networks.

3. Experimental Observations

In this section, we first present a brief overview of the ingredients needed to perform the experiments such as database, protocol, and evaluation metrics.

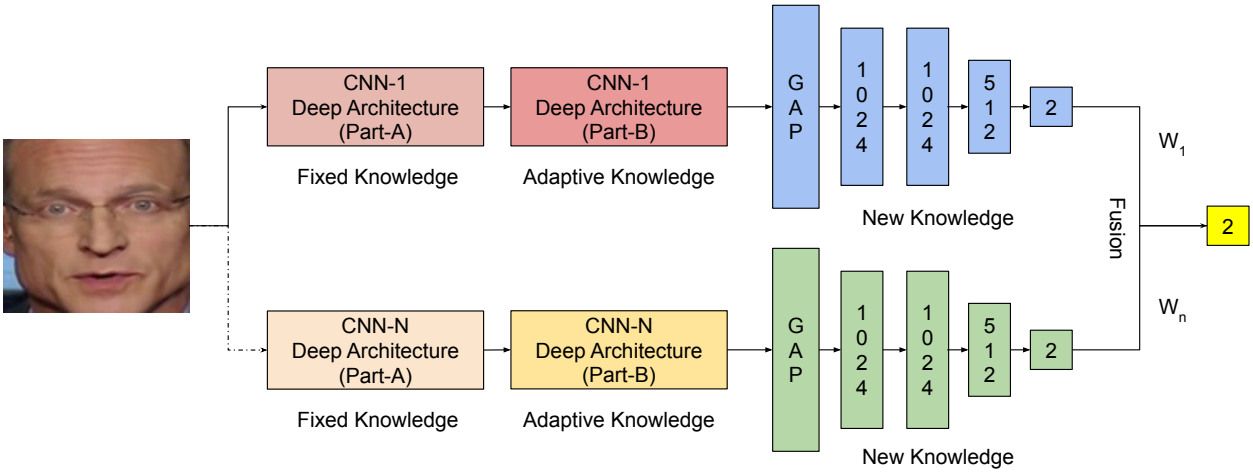


Figure 2. Visualization of the proposed multi-level and multi-branch deepfake detection architecture. The proposed architecture consists of 'N' deep CNN architecture divided into fixed base knowledge and adaptive knowledge. In both networks' novel layers which are bias-free from any specific set of training images layers are added and initialized with random weights. In the end, decision fusion has been performed to achieve better performance. GAP refers to the global average pooling layer. w_1 and w_n (where $w_1 + \cdot\cdot + w_n = 1$) are weights for fusion.

3.1. Experimental Ingredients

The strength of the proposed multi-branch and multi-point deepfake detection architecture is extensively evaluated using three different datasets namely Face Forensics++ (FF++) [45], Celeb-DF2 [32], and Deepfakes [15]. The deepfake subset of FF++ contains 1000 videos comprising 720 videos for training and 140 videos for validation and testing. The dataset comes with three quality variants, in which we have used C23 (HQ) and C40 (LQ) videos for comparisons with the existing works. The Celeb-DF2 is the high-quality variant of deepfake attack compared to FF++ aims to reflect the high-quality videos surfacing over the internet. The dataset contains 590 real and 5639 deepfake videos. Ciftci et al. [15] have introduced another diverse deepfake dataset which consists of 142 high-quality real and deepfake videos. The results are reported using the pre-defined protocols and evaluation metrics used in the existing papers to make direct and fair comparisons. For instance, intra-dataset training testing on FF++ is performed using the pre-defined train-test split of the individual compression quality has been performed. In the cross-datasets, the FF++-trained model is tested on Celeb-DF2. Apart from these datasets, we have also collected images from the *real-world environment for an open-set evaluation*.

3.2. Results and Analysis

First, we present the results and comparative analysis on the FF++ dataset [45] followed by the results on the deepfakes dataset [15]. In the end, the generalizability and robustness of the proposed and existing algorithms are tested on the

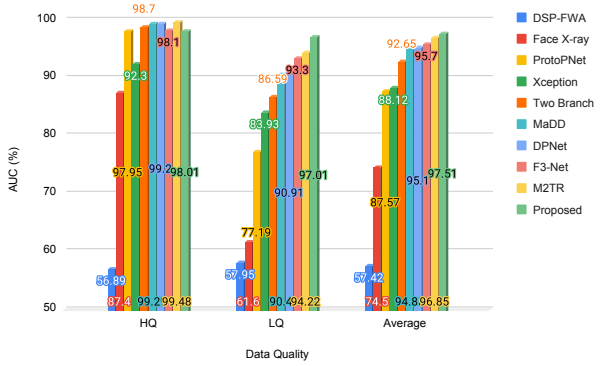


Figure 3. Quantitative frame-level detection results on the FF++ dataset under medium compression (HQ) and high compression (LQ). Average represents the average performance of each algorithm obtained on both the LQ and HQ datasets. The results are compared with state-of-the-art algorithms. The proposed algorithm yields state-of-the-art performance on low-quality (LQ) images where each algorithm suffers severe degradation and showcases its limited generalizability.

Celeb-DF2 dataset [32] which is used for unseen database training-testing. The quantitative results and comparison of the proposed algorithm on two different quality subsets of the FF++ dataset are reported in Fig. 3. The comparisons have been performed using several complex state-of-the-art and recent algorithms to demonstrate the effectiveness of the proposed algorithm. The algorithms chosen for comparison are: DSP-FWA [29], Face X-ray [28], ProtoPNet [12], Xception [45], Two Branch [37], MaDD [61], DPNet

[53], F3-Net [44], and M2TR [55]. It is generally seen that the deepfake detection algorithms do not generalize well against different image qualities which here in the dataset refers to different compress rates. For instance, the existing state-of-the-art algorithm M2TR [55] yields more than 99 AUC value when it is evaluated on the high-quality subset of the dataset. Whereas, the algorithms see a significant drop of more than 5% when tested on low quality or high compression (LQ) subset of the dataset. The proposed algorithm which yields 98.01% AUC value on the high-quality or low compression (HQ) subset shows a marginal drop (0.5%) in the AUC value when it is tested on the low-quality subset of the dataset. It shows that the proposed algorithm is not only able to detect the deepfake images/videos effectively but is generalized against different image qualities. The generalizability can be easily checked through the average AUC computed from the results on HQ and LQ subsets of the dataset. The proposed algorithm can outperform several recent algorithms such as MaDD [61] and DPNet [53] which are highly computationally expensive and lack generalizability. The generalizability of the deepfake detection algorithms is a serious concern and needs proper attention [36, 57]. Our proposed simple fusion algorithm puts a strong step toward that goal by yielding thrilling results and shows an exciting direction to achieve generalizability at a lower computational overhead.

Apart from existing state-of-the-art deepfake detection algorithms, we have also compared the performance of the proposed algorithm with recent deep architecture namely vision transformers (ViT)[1]. We have used two architectures termed ViT-v1 [17] and ViT-v2 [27]. The ViT-v1 and ViT-v2 yield 85.43% and 80.18% accuracy of high-quality (HQ) deepfake videos, respectively, which is at least 8.58% lower than the proposed algorithm. On the LQ videos of the FF++ dataset, the performance of ViT-v1 and ViT-v2 is 9.40% and 11.97% lower than the proposed algorithm, respectively. We assert that the superior performance of the proposed simple fusion and knowledge-enhanced network than these complex networks pave the way for a revisit to the fundamentals of the image classifiers to develop simple and effective classifiers. The prime reason for our effective deepfake identification study contrary to literature is the fact that the deepfake detection research is highly biased [5, 45, 65] towards the utilization of Xception and ignored the potential of other networks such as VGG and MobileNet.

Another dataset that is used for experimental evaluation is the Deepfakes dataset [15] and the experimental results along with comparison are reported in Table 1. Contrary to the FF++ dataset, the Deepfakes dataset contains high-resolution and high-quality deepfake videos/images. The reason for the generation of high-quality as explained by the authors is to reflect the real-world videos that exist on social media websites and are of high quality. The authors claim that the detection of high-quality deepfake videos is hard and this makes their dataset challenging compared to other datasets. Similar to the FF++ dataset, the comparison on the deepfake dataset has been performed with several benchmark algorithms: InceptionV3 [49], Xception [13], ConvLSTM [56], Emsemble [51], and Fake Catcher [15]. The existing algorithms range from the training of straightforward deep CNNs to the development of sophisticated classifiers such as Fake Catcher. The proposed algorithm achieves state-of-the-art (SOTA) deepfake detection performance on another challenging Deepfakes dataset. In comparison to the SOTA works the performance of the proposed algorithm is at least 3.64% better when only the face images are used for testing. Even when the videos that contain an ensemble of information from different face regions are used for evaluation, the proposed algorithm outperforms the SOTA algorithm by 3.71%. On the previous dataset, through the experimental evaluation, we observed that the proposed algorithm is agnostic to compression effect and achieves SOTA performance. Now, here through the experiments on a high-quality deepfake dataset, it is established that the proposed algorithm surpasses multiple existing algorithms by a large margin and is agnostic to image quality.

Algorithm	Face ↑	Video ↑
Simple CNN	54.56	48.88
InceptionV3 [49]	60.96	68.88
Xception [13]	56.11	75.55
ConvLSTM [56]	44.82	48.83
V1 [51]	–	82.22
V3 [51]	–	73.33
Emsemble [51]	–	80.00
Fake Catcher [15]	87.62	91.07
Proposed	**91.26**	**94.78**

Table 1. Deepfakes database [15] results. Comparison of the proposed image engineering enhanced attention network with the several networks in terms of detection accuracy (%). The proposed algorithm yields almost perfect deepfake detection performance and achieves existing state-of-the-art algorithms. The existing results are taken from [15].

In the literature, another experimental setup has been used to evaluate the strength of the deepfake detection algorithm i.e., to perform the testing on an entirely unseen dataset which does not used in the training. The dataset that is extensively used for that purpose is Celeb-DF2 [32]. Table 2 shows the performance of the proposed algorithm and several existing algorithms when trained on the FF++ dataset and tested on FF+ and Celeb-DF2. The results showcase how the accuracy of each existing algorithm suffers due to variations in the distribution of the testing set.

[1]We want to mention here that these ViT are trained from scratch on the training set of the FF++ dataset only.

Method	FF++	Celeb-DF2
Mesolnception4 [2]	0.8300	0.5360
HeadPose [58]	0.4730	0.5460
FWA [29]	0.8010	0.5690
VA-MLP [38]	0.6640	0.5500
Xception-raw [45]	0.9970	0.4820
Xception-c23 [45]	0.9970	0.6530
Xception-c40 [45]	0.9550	0.6550
Multi-task [41]	0.7630	0.5430
Capsule [42]	0.9660	0.5750
DSP-FWA [29]	0.9300	0.6460
F^3-Net [44]	0.9797	0.6517
Two-Branch [37]	<u>0.9318</u>	**0.7341**
EfficientNet-B4 [50]	0.9970	0.6429
Nirkin et al. [43]	0.9900	0.6600
Multi-Attention [61]	0.9980	0.6744
DPNet [53]	0.9920	0.6820
M2TR [55]	0.9950	0.6570
MD-CSDNetwork [5]	0.9970	0.6877
Proposed	0.9801	0.7035

Table 2. Cross-dataset evaluation (AUC) on Celeb-DF2 [32]. The model is trained on FF++ and tested on the Celeb-DF dataset. The results in the first column report the AUC values when tested only on the deepfake class in FF++. Our method outperforms most of the listed methods in cross-generalization.

It is seen from the quantitative results that the majority of the recent algorithms have achieved almost perfect AUC when trained and tested on the same FF++ dataset. However, the performance of each algorithm drops significantly when evaluated on the unseen Celeb-DF2 dataset. For instance, the recent multi-attention algorithm [61] which achieves 99.80% AUC value on the seen dataset testing, reported a significant drop and yields only 67.44% AUC on the unseen dataset testing set. The proposed algorithm shows slightly lower performance on the seen dataset testing images but achieves a significantly higher generalizability score when tested on the unseen dataset testing images. The best-generalized algorithm is the *Two branch algorithm* which achieves the AUC value of 73.41% on the Celeb-DF2 dataset. However, the performance of the two-branch is 4.83% lower than the proposed algorithm on the seen dataset testing images. It shows that the proposed algorithm is not only effective but also generalized in handling the images whether coming from seen or unseen dataset testing images and agnostic to the image/video quality. Fig. 4 shows the average AUC performance of the existing SOTA and the proposed energy-efficient algorithms. The evaluation has been done on the testing set of FF++ and Celeb-DF2 while the algorithms are trained on the FF++ dataset. *The proposed algorithm shows the highest average AUC value and surpasses several computationally hungry existing algorithms in challenging open-set evaluation settings.*

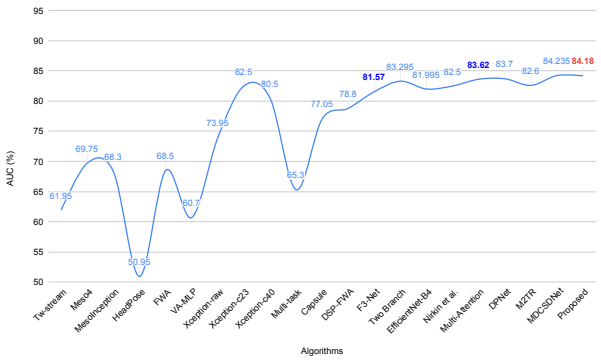


Figure 4. Average AUC (%) performance when the deepfake detection algorithms that are trained on the training set of the FF++ dataset and tested on the test set of FF++ and Celeb-DF-2. The top three values are bold and colored and the proposed algorithm reported the highest value. (Best viewed in color)

CNN-1		CNN-2		HQ	LQ
Name	BP	Name	BP		
VGG-16	10	Inception-V3	10	0.9715	0.9534
VGG-16	10	Xception	10	0.9745	0.9682
VGG-16	10	DenseNet-121	110	0.9801	0.9701
VGG-16	10	MobileNet	12	0.9755	0.9529
Inception-V3	10	DenseNet-121	110	0.9278	0.9406

Table 3. Ablation study of the deepfake detection task on the different quality of FF++ when varying CNNs are trained with their corresponding breakpoints (BP). The results are reported in terms of AUC.

3.3. Ablation Studies

In this section, different ablation study results are reported. For instance, the performance of the proposed algorithm when different types of deep architectures are used and their broken points are adaptively set to segregate the features into fixed and adaptive. Later, computational complexity or the time taken by the proposed algorithm in its training and testing is also reported to reflect the sustainability strength helpful in deploying the algorithm on mobile devices.

CNN Architectures: The results reported earlier using the proposed algorithms are computed when VGG-16 [47], MobileNet [21], and Xception [13] models are used for the model training and fused for evaluation. Each model is broken down at layer 10, the layers before breakpoints are kept fixed with the assumption that they consist of generalized image features as mentioned above, and to learn the task-specific features the future layers are made adaptive. After that few extra layers are added that do not have any previous knowledge of any image recognition task, i.e., they are

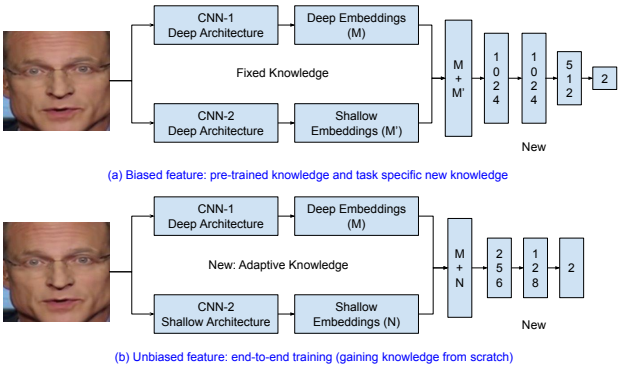


Figure 5. Different deepfake evolved architecture based on what knowledge they already persist and what they acquire for the task at hand.

unbiased and are randomly initialized and trained for only deepfake detection. Each network containing fixed, adaptive, and new layers is trained independently, in the end, their decision probabilities are combined for the final decision. In the first ablation study, we have used other popular deep CNNs including DenseNet [22] and InceptionV3 [49] to see their impact on deepfake detection. The quantitative comparison with different CNNs on the different quality subsets of the FF++ dataset is reported in Table 3. Through the results, we have found that the combination of VGG, MobileNet, and Xception outperforms the combination of other networks. Apart from using the breakpoints (BP) shown in Table 3, we have also studied several other BP for different networks; however, empirically found that these BP as the best points concerning computational cost and detection performance. For instance, we have also evaluated the different BPs of the VGG network to layers 6 and 3; however, no significant improvement in accuracy is observed but the finetuning cost of the network increased. Similarly, for DenseNet, the different BPs varied from layer 90 to layer 60 are applied. Due to the trade-off between computational cost and accuracy improvement, in this research, we have set the BP shown in Table 3.

Biased and Unbiased Network Feature Fusion Detectors: In this study, we have performed the experiments using two different architectures (shown in Fig. 5) other than the proposed one which utilizes the score fusion strategy. In the first case, i.e., biased knowledge, the feature fusion strategy is adopted; however, the networks might be biased due to their pre-training on the ImageNet dataset, and a few new layers are added to learn task (deepfake detection) specific knowledge. The second architecture (termed unbiased network) also utilizes the feature fusion strategy; however, now both the networks are randomly initialized and trained end-to-end for deepfake detection. The first architecture yields an AUC value of 0.7353 using VGG16

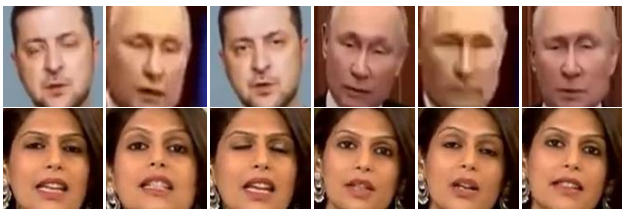


Figure 6. Real-World disinformation deepfake videos from the Ukraine-Russia War. The first row is deepfake and the second row is real images.

		Predicted →		
		Real	DF	Total
True ↓	Real	**639**	42	681
	DF	17	**664**	681
	Total	656	706	**1362**

Table 4. Confusion matrix of the real-world disinformative deepfake examples detection using the proposed fast and generalized deepfake detection algorithm.

and DenseNet121. Whereas, the second architecture utilizes one shallow (5 conv layers) and one deep (11 conv layers) and achieves 0.9275 AUC on the C40 set of the FF++ dataset. The experiment shows that the finetuning of features and pre-training of a network can have a huge impact on the downstream task. We want to mention that, the proposed algorithm which consists of all three: biased, unbiased, and task-specific new features achieves the highest AUC of value 0.9701 and shows the importance of contrasting knowledge.

Computational Time: The proposed algorithm with VGG, MobileNet, and XceptionNet as base architectures is trained on a single RTX 2080 GPU machine. The training of the proposed algorithm took approximately 50 minutes on the FF++ dataset and achieved state-of-the-art performance on several datasets. On top of that, it is found generalized against several data variants such as compression quality and image quality. We also want to highlight that for the training, a limited number (10 only) of frames/faces are randomly selected from each training video given in the datasets. It shows that the proposed algorithm is neither computationally hungry nor data-hungry. We want to mention that the evaluation has been performed on the standard test set of each dataset.

3.4. Real-World Disinformation Prevention

While the proposed algorithm has shown exciting results on the existing research dataset; however, the question we raise is whether can it detect the deepfake videos spreading over the Internet mediums such as YouTube[2]? The recent

[2]https://www.youtube.com/watch?v=pfsdvbacYac

Figure 7. Grad-CAM [46] heatmap visualization reported on unconstrained images collected as part of our evaluation dataset. The heatmap information is computed using the three CNN architectures used for the development of a deepfake detection algorithm.

surge of deepfake videos on the Russian-Ukraine conflict including the disinformation spread due to fake videos of both Russia and Ukraine's president shocked the world and raised the question of whether one can trust digital media blindly. Therefore, to protect the digital media from false news and also to further test the robustness of the proposed algorithm in an open-set setting, we have collected 681 real and 681 deepfake images captured in unconstrained settings, i.e., varying in terms of poses, expression, and image quality. A few samples of the real-world disinformative deepfake samples are shown in Fig. 6. No special preprocessing has been applied to the collected images, only the face regions are cropped using the Viola-Jones face detector [54] and resized to a fixed size to provide them as input to the classifier. The proposed score fusion algorithm which is trained on the FF++ dataset evaluated on the collected real-world yields $\mathbf{95.68}\%$ deepfake detection accuracy. In terms of AUC, the proposed algorithm achieves a value of over 98%, and the confusion matrix of the evaluation is also provided in Table 4. The confusion matrix shows that the proposed algorithm while the proposed algorithm is effective and is *not biased toward any class. In other words, the proposed algorithm yields higher accuracy on both classes*. The evaluation of the real-world videos/images and the high performance even when those images are not seen at the time of training a classifier shows that the proposed algorithm which is computationally light might be one of the potential options to prevent the spread of deepfake. The Grad-CAM [46] heatmap visualization shown in Fig. 7 can be interpreted in the following manner: (i) the different network focuses on the different regions of faces and hence their combination yields better performance, (ii) VGG architecture strongly focuses on high-frequency regions such as mouth, eyes, and nose, and (iii) MobileNet puts its focus on entire face region while making a decision.

4. Conclusion

In recent times we have witnessed the use of deepfake videos for a variety of purposes ranging from personal revenge to political advantages to monetary benefits. On top of that, the existence of these videos is not limited to any particular social media application and can easily be found without restrictions on every possible social networking platform. Due to their existence on different platforms and due to varying electronic screen characteristics, deepfake videos might suffer in terms of compression effects and image quality. By looking at the stealthy purpose of these videos their detection is critical and one important strength of the detection algorithm should be the generalizability to handle these variations. Another important factor that needs to be tackled is the deployment of the algorithm on computationally limited devices including mobile phones. Therefore, keeping all this in mind, in this research, we have proposed a multi-branch and multi-level deepfake detection algorithm. The proposed algorithm consists of the concept of knowledge breakout where the generic features are kept intact and adaptive knowledge is finetuned for the downstream task. The proposed algorithm is evaluated in several challenging and generalized conditions on several databases to demonstrate its effectiveness in deepfake detection. The proposed algorithm surpasses several existing algorithms in these challenging scenarios by significant margins. In the future, similar to countering the video deepfake detection in a computationally efficient fashion we aim to develop deepfake detection to counter the audio and multi-modal deepfake contents.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In {*USENIX*}

symposium on operating systems design and implementation ({OSDI} 16), pages 265–283, 2016. 3

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018. 6

[3] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE International Joint Conference on Biometrics*, pages 659–665, 2017. 1

[4] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini K Ratha. Image transformation based defense against adversarial perturbation on deep learning models. *IEEE Transactions on Dependable and Secure Computing*, 2020. 2

[5] Aayushi Agarwal, Akshay Agarwal, Sayan Sinha, Mayank Vatsa, and Richa Singh. MD-CSDNetwork: Multi-domain cross stitched network for deepfake detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 2, 5, 6

[6] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Afzel Noore. MagNet: Detecting digital presentation attacks on face recognition. *Frontiers in Artificial Intelligence*, 4: 643424, 2021. 1

[7] Akshay Agarwal, Afzel Noore, Mayank Vatsa, and Richa Singh. Generalized contact lens iris presentation attack detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):373–385, 2022. 3

[8] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Crafting adversarial perturbations via transformed image component swapping. *IEEE Transactions on Image Processing*, 31:7338–7349, 2022. 1

[9] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2814–2822, 2020. 1

[10] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74, 2021. 3

[11] Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *IEEE International Conference on Pattern Recognition*, pages 5012–5019, 2021. 2

[12] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *arXiv preprint arXiv:1806.10574*, 2018. 4

[13] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 5, 6

[14] François Chollet et al. Keras. https://keras.io, 2015. 3

[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. 1, 4, 5

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[18] Mengnan Du, Shiva K. Pentyala, Yuening Li, and Xia Hu. Towards generalizable deepfake detection with locality-aware autoencoder. 2020. 2

[19] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pages 3247–3258. PMLR, 2020. 2

[20] Mehak Gupta, Vishal Singh, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Generalized iris presentation attack detection algorithm under cross-database settings. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5318–5325, 2021. 3

[21] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 6

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3, 7

[23] Rohit Keshari, Mayank Vatsa, Richa Singh, and Afzel Noore. Learning structure and strength of cnn filters for small sample size training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9349–9358, 2018. 3

[24] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S Woo. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *1st workshop on synthetic multimedia-audiovisual deepfake generation and detection*, pages 7–15, 2021. 2

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[26] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. *arXiv preprint arXiv:2103.10094*, 2021. 1

[27] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021. 5

[28] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 4

[29] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1, 4, 6

[30] 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. 1

[31] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition*, 2020. 1

[32] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 4, 5, 6

[33] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[34] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017. 3

[35] Puspita Majumdar, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Evading face recognition via partial tampering of faces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–20, 2019. 1, 2

[36] Puspita Majumdar, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Facial retouching and alteration detection. In *Handbook of Digital Face Manipulation and Detection*, pages 367–387. Springer, Cham, 2022. 5

[37] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*, pages 667–684, 2020. 4, 6

[38] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Applications of Computer Vision Workshops*, pages 83–92, 2019. 6

[39] Aman Mehra, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Motion magnified 3-d residual-in-dense network for deepfake detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1):39–52, 2023. 1

[40] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021. 2

[41] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–8, 2019. 6

[42] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2307–2311, 2019. 6

[43] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 6

[44] 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. 5, 6

[45] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 1, 4, 5, 6

[46] 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 *IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 6

[48] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *AAAI Conference on Artificial Intelligence*, pages 13583–13589, 2020. 2

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3, 5, 7

[50] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 6

[51] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Detecting both machine and human created fake face images in the wild. In *international workshop on multimedia privacy and security*, pages 81–87, 2018. 5

[52] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 2

[53] Loc Trinh, Michael Tsang, Sirisha Rambhatla, and Yan Liu. Interpretable and trustworthy deepfake detection via dynamic prototypes. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1973–1983, 2021. 5, 6

[54] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2): 137–154, 2004. 8

[55] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. M2tr: Multi-modal multi-scale transformers for deep-

fake detection. *arXiv preprint arXiv:2104.09770*, 2021. 2, 5, 6

[56] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 5

[57] Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389, 2022. 5

[58] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8261–8265, 2019. 1, 6

[59] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 3

[60] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3

[61] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021. 2, 4, 5, 6

[62] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021. 2

[63] Zheng Zhao, Penghui Wang, and Wei Lu. Detecting deepfake video by learning two-level features with two-stream convolutional neural network. In *International Conference on Computing and Artificial Intelligence*, pages 291–297, 2020. 2

[64] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021. 2

[65] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM international conference on multimedia*, pages 2382–2390, 2020. 5