Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection

Ying Xu Kiran Raja Marius Pedersen
Norwegian University of Science and Technology (NTNU), Norway
{ying.xu; kiran.raja; marius.pedersen} @ntnu.no

Abstract

DeepFakes detection approaches have to be agnostic across generation type, quality, and appearance to provide a generalizable DeepFakes detector. Limited generalizability will hinder wide-scale deployment of detectors if they cannot handle unseen attacks in an open set scenario. We propose a generalizable detection model that can detect novel and unknown/unseen DeepFakes using a supervised contrastive (SupCon) loss. As DeepFakes can resemble the original image/video to a greater extent in terms of appearance and it becomes challenging to secern them, we propose to exploit the contrasts in the representation space to learn a generalizable detector. We further investigate the features learnt from our proposed approach for explainability. The analysis for explainability of the models advocates the need for fusion and motivated by this, we fuse the scores from the proposed SupCon model and the Xception network to exploit the variability from different architectures. The proposed model consistently performs better compared to the single model on both known data and unknown attacks consistently in a seen data setting and an unseen data setting, with generalizability and explainability as a basis. We obtain the highest accuracy of 78.74% using proposed SupCon model and an accuracy of 83.99% with proposed fusion in a true open-set evaluation scenario where the test class is unknown at the training phase. The paper also aligns with reproducible research by making the code available [1].

1. Introduction

DeepFakes [2] can be used for a variety of purposes, such as pure entertainment applications to targeted identity attacks [5, 61, 64]. However, the barriers to the creation of convincing DeepFakes in terms of required computer capabilities, specialised knowledge, and other resources, such as training data, are constantly decreasing, making successful use of disinformation attacks ever more likely.

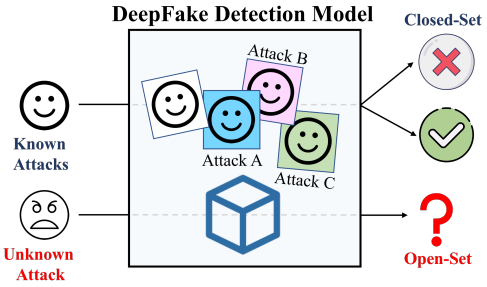

Figure 1: Illustration of DeepFakes detection models as open-set problem.

Face manipulation techniques have evolved from approaches needing manual effort (for instance, Adobe PhotoShop) to approaches needing minimal training [3, 1]. The recent advancements in Generative Adversarial Networks (GANs) along with the availability of GPUs have now led to the creation of completely artificial yet hyper-realistic content which can highly challenge human observers [70, 40, 65, 29, 30, 8, 10]. Complementing the approaches for generation, a number of approaches for detecting DeepFakes have been proposed exploiting artifacts, inconsistencies in images and discontinuity in video [37, 44, 17, 16, 25, 50, 31, 21, 24, 13, 51, 39].

Despite the high quality of DeepFakes, with careful scrutiny one can observe that the generated images and videos present certain artifacts that can help in detecting the manipulated content. A number of DeepFakes detection approaches have been proposed using artifact clues from DeepFakes [37, 44, 17, 16, 25, 50]. Many approaches rely on looking at inconsistencies of videos [31, 21], change of temporal and spatial information [54], frequency information [24, 13, 51, 39] or audio inconsistencies [47, 34]. Most of these works extract the features using either handcrafted mechanism (for instance, texture features like Local Ternary Patterns (LTP) [58], Local Binary Patterns

[1] https://github.com/xuyingzhongguo/deepfake_supcon

[2] DeepFakes in this article refers to family of different fake creation approaches such as FaceSwap (FS) [3], DeepFakes (DF) [2], NeuralTextures (NT) [62] and Face2Face (F2F) [63].

(LBP) [49], Scale-Invariant Feature Transform (SIFT) [41]) or deep features (Convolutional Neural Network (CNN), Recurrent Neural Network (RNN)), GAN [28]) followed by learning a classifier.

DeepFakes detection can be posed as a closed set problem where the training and testing data are drawn from the same label and feature spaces, for instance, detecting images created using FaceSwap [3] based on training with the same. However, given the rapid progress of generation approaches, the detectors should be equipped to detect unknown/unseen testing data which can emerge from different labels and feature spaces. In the context of DeepFakes detection, this can be parallel to detecting FaceSwap when the detector is trained on Face2Face, making it an open set problem. DeepFakes detection approaches have focused heavily on detecting and classifying known attack types in closed-set classification. However, newer DeepFakes generation mechanisms make the detection algorithms unreliable and non-generalizable by degrading the performance of the detector [27, 6] as no exception to machine learning based classifiers. A simple illustration of this problem is shown in Figure 1. The reasons behind the failure of detection models towards unseen content can, to some extent, be attributed to different generation principles, which often result in different data distributions, feature spaces, and appearance properties of images or videos.

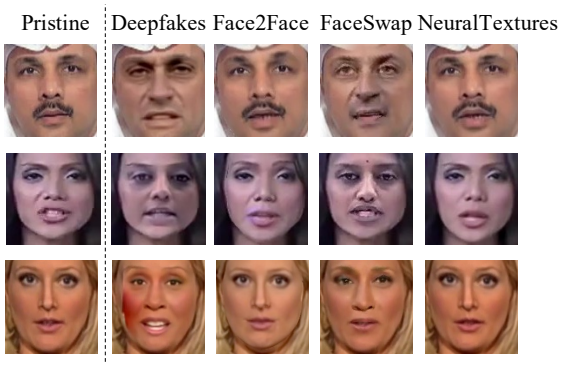

Figure 2: Sample face images from different attacks of FaceForensics++ dataset used in this work. The first column presents the pristine (non-manipulated) frames and four other columns present from DeepFakes (DT), Face2Face (F2F), FaceSwap (FS) and NeuralTextures (NT) respectively.

Posing DeepFakes detection as an open set problem and noting limited works in this direction, we assert the need for a generalizable detection approach for making reliable decisions on unknown/unseen generation types in addition to known/seen generation types. While we note that the manipulation type can influence the feature space based on different approaches, we also assert that the real images exhibit a different feature space, which often can be trusted to design a detector [3]. Thus, we propose to exploit the contrasts between non-manipulated images against a set of manipulated images in an attempt to generalize the detector towards unknown manipulation types or unseen data. We specifically employ Supervised Contrastive (SupCon) Learning to guide the detector to classify the non-manipulated images efficiently and at the same time differentiate the manipulated images using FaceForensics++ dataset [53] consisting of four different manipulation types such as DeepFakes (DF) [2], FaceSwap (FS) [3], Face2Face (F2F) [63] and NeuralTextures (NT)[62]. Examples taken from this dataset are shown in Figure 2. The proposed approach is, in addition, benchmarked against four different state-of-the-art models such as Convolutional LSTM based residual network - CLRNet [60], Transfer learning-based Autoencoder with Residuals (TAR) [35], Generalized Zero and Few-Shot Transfer approach [7] and Xception network [15]. Further, analyzing the features learnt from each proposed approach, we look at explaining the efficiency of the model. Trying to explain the efficiency of the two top-performing models, we also discover the complementarity of the features, suggesting a solid motivation for fusion. We, therefore, make three key contributions:

- **New Approach**: A new approach exploiting the contrastive feature space between non-manipulated and manipulated images is proposed for DeepFakes detection. To the best of our knowledge, the proposed Supervised Contrastive (SupCon) learning is the first work exploring the idea of contrastive learning tested on the publicly available FaceForensics++ dataset consisting of four different types of attacks.
- **Generalizability**: Considering the generalization aspect of the proposed approach, we report the results in a true open-set scenario by using three known manipulation types for training and one unknown type for testing. The proposed approach is evaluated on all combinations of known-training and unknown-testing sets, where the results positively affirm the proposed approach.
- **Explainability**: In an attempt to explain the performance obtained from the model, we analyze the features from a visual aspect using the Heatmap visualization and Uniform Manifold Approximation and Projection (UMAP) both of which corroborate the initial assertion. Motivated by the observations, we identify the complementarity of our proposed model with another top-performing model based on Xception. We resort to a weighted score level fusion of the proposed SupCon and Xception models to provide a generalizable and explainable model.

[3]The feature space may vary across spatial resolutions and on other factors such as noise, capture conditions. We treat this as out of scope in this work as we focus on publicly available datasets.

In the rest of the paper, we present a set of related works in Section 2 and a brief explanation of contrastive learning in Section 2.3. The proposed approach is further detailed in Section 3 and 5 along with the rationale for the approach. We provide an analysis on explainability in Section 4 with the set of experiments and results on generalizability detailed in Section 6. Towards the end of the article, we present the limitations of the current work and conclusions with potential future works in Section 7 and 8, respectively.

2. Related Works

2.1. DeepFakes Detection

Several types of deep networks have been used for DeepFakes detection over the last few years. A general approach is to detect visible artifacts in the image or video of the face, and the methods following this approach highlight specific failures in the generation process that does not faithfully reproduce real face details helping in detection [37, 44, 38, 36, 45, 38, 52]. A number of papers have employed Convolutional Neural Networks (CNNs) based methods for detecting such artifacts [42, 43, 68, 53, 23, 48, 22, 9]. Another set of approaches were specifically designed explicitly taking into account the temporal directional changes in videos [31, 54, 21]. Two-stream networks [71, 11, 56, 66], combining two different kinds of features for detection and classification tasks, are also gaining more popularity on this topic. Other focuses have been put on the frequency domain [24, 14, 51], as well as GAN fingerprints [70, 40, 65, 29, 30, 8].

2.2. Limitations in Generalization of DeepFakes Detection

While a number of works are proposed for detecting DeepFakes, the most noted works in the previous section correspond to closed-set experiments where the training and testing set corresponded. Minimal works have tried to address the problem of generalization of DeepFakes detection [7, 60, 59, 33, 35]. These works have focused on domain adaptation and transfer learning to minimize the task of learning parameters in an end-to-end manner [7, 60, 59, 33, 35]. Cozzolino *et al*. [18] proposed ForensicTransfer where the generalization aspect was studied using a single detection method for multiple target domains. Transfer learning-based Autoencoder with Residuals (TAR) [35] proposed recently employs the residuals from auto-encoders to address the generalizability. Despite the limited works, we note the best accuracy in an open-set testing scenario results in 82.73%, 64.69%, 49.74% and 55.59 using Xception network for DF, F2F, FS, and FT while TAR achieves 75.25%, 72.90% and 51.65% on DF, F2F, FS. However, the experimental protocols in Lee *et al*. [35] do not consider all subsets of DF, F2F, FS, and FT in training and testing sets, making it not a pure open-set problem.

2.3. Contrastive Learning and DeepFakes Detection

Contrastive learning focuses on learning the common features between instances of the same kind and distinguishing the differences between different types of samples. Contrastive learning only needs to learn to distinguish the data in the feature space of the abstract semantic level, making the models attractive. Fung *et al*. [26] used unsupervised contrastive learning to address DeepFakes detection as an attempt towards generalization across datasets. However, Fung *et al*. [26] did not fully exploit the label information following an unsupervised setting. We, however, note that using class information to determine whether images belong to the same category is crucial to make the features of similar pictures close to each other. With this in the backdrop, we assert the usefulness of supervised contrastive loss [32] as a suitable approach for the generalization task. To the best of our knowledge, we explore Supervised Contrastive (SupCon) learning for the first time to address DeepFakes detection using Efficient-B0 [57] as the backbone and simultaneously achieve generalization.

3. Generalizable DeepFakes Detection using SupCon

The key idea of our method is to employ Supervised Contrastive (SupCon) learning to solve the generalization problem of DeepFakes detection. We first provide the fundamentals of Contrastive Learning for the convenience of the reader before providing the details of the proposed approach as illustrated in Figure 3.

3.1. Contrastive Learning

For data point x, the goal of contrastive learning is to learn features $f(x)$ such that

$$score(f(x), f(x^+)) >> score(f(x), f(x^-)) \quad (1)$$

where x^+ is a data point similar to x, referred to as a positive sample and x^- is a data point dissimilar to x, referred to as a negative sample.

The fundamental working of contrastive learning can be formulated as a score function. This $score$ is a metric that measures the similarity between two features. A simple framework for contrastive learning of visual representations (SimCLR) [12] was proposed to build a self-supervised contrastive model.

This contrastive learning strategy suits our goal of the generalization problem to discriminate DeepFakes from non-manipulated real visual media (images/videos). However, the learning strategy in Equation 1 cannot be directly employed as it does not consider the category information

to determine whether images belong to the same class. We, therefore, reformulate the loss in a supervised manner rather than a self-supervised approach. Seeking a suitable solution, we adopt supervised contrastive (SupCon) loss [32] in our proposed approach, considering the class/category/label information.

3.2. Learning a Classifier for DeepFakes Detection

Given an input batch N of data x in $X \in R^D$, data augmentation is applied twice to obtain two identical copies of the batch. We employ four different implementations of the data augmentation module: AutoAugment [19], RandAugment [20], SimAugment [12], and Stacked RandAugment [67] as recommended in earlier work [32] to make the learning robust and invariant to noisy data. Both copies are further forward propagated through the encoder $f(\cdot)$ to obtain a normalized embedding. During training, this intermediate representation h is further propagated through a projection network $g(\cdot)$ that would be discarded at inference time. The supervised contrastive loss is computed on the outputs of the projection network. The learnt representation from the output of the network is further used to learn a classifier. Furthermore, we employ Efficient-B0 [57] for

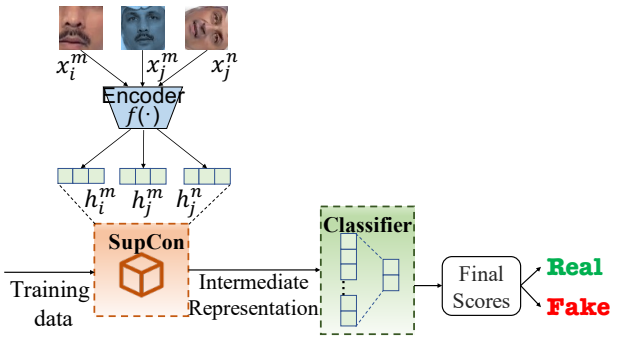

Figure 3: Framework for proposed approach using SupCon for DeepFakes detection. The intermediate representation of each input data produced through SupCon is further fed into a classifier. The score obtained from the classifier is then used to determine the pristine nature of image.

the encoder network $f(\cdot)$ to significantly reduce the overhead of computation and yet retain the performance benefits. The normalized activations of the final pooling layer are used as the representation features for our proposed approach.

We further train a linear classifier on the learned representations using a cross-entropy loss that can differentiate pristine/non-manipulated images against manipulated images. For a set of N randomly sampled sample and label pairs, $\{x_\ell, y_\ell\}_{k=1...N}$, the corresponding batch used for training consists of $2N$ pairs, $\{\tilde{x}_\ell, \tilde{y}_\ell\}_{k=1...2N}$, where $\tilde{x}_{2k}$ and $\tilde{x}_{2k-1}$ are two random augmentations of $x_k (k = 1...N)$ and $\tilde{y}_{2k} = \tilde{y}_{2k-1} = y_k$. The loss function for supervised contrastive learning is defined as:

$$\mathcal{L}^{sup} = \sum_{i=1}^{2N} \mathcal{L}_i^{sup}$$

$$\mathcal{L}_i^{sup} = \frac{-1}{2N_{\tilde{y}_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{[i \neq j]} \cdot \mathbb{1}_{[\tilde{y}_i = \tilde{y}_j]} \cdot \tag{2}$$

$$log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

The supervised contrastive loss is an enlargement of the self-supervised contrastive loss as seen from Equation 2 where the supervised contrastive loss expands the number of positive pairs of $\tilde{x}_i$ such that all sub-data with the same label are regarded as positive pairs. Equation 2 calculates the weighted average of the similarities between $\tilde{x}_i$ and all positive pairs.

4. Explainable Analysis of DeepFakes Detection : What makes SupCon Usable?

We analyse the proposed approach founded on theoretical assertions for explainability. We provide the analysis on the basis of two different approaches where the first analysis is conducted using heatmaps [55] and second analysis is conducted using Uniform Manifold Approximation and Projection (UMAP) [46]. While heatmaps allow a better visualization of what has been learned by the network, UMAP provides topology explanations of the learnt features.

Figure 4 presents the heatmaps corresponding to the last layer of the proposed SupCon overlaid with different kinds of DeepFakes (DF, F2F, FS, and NT). We note that the SupCon focuses on the silhouette on the face region where the manipulation exists. Our assertion of the SupCon applicable to DeepFakes detection is corroborated through visual analysis. Furthermore, we also do a similar analysis on the Xception network as presented in Figure 4. Noting from the visual analysis, it is evident that the Xception network focuses on the regions inside the face region such as foreheads, eyebrows, and eyes. The focus of heatmaps on different areas of the face region clearly suggests the complementarity of the networks. The analysis thus forms the basis for our proposed approach of fusion as explained further in Section 5.

We further present UMAP analysis in Figure 5 using the features extracted from the penultimate layer (before the classifier). We employ a model trained using pristine images and images from F2F, FS, and NT to conduct this analysis. The purple dots correspond to the manipulated images and the red ones correspond to the pristine images in Figure 5. As noted from this analysis, the features from SupCon are mixed and spread across, while the UMAP for Xception indicates clear boundaries between pristine and manipulated images. We further concatenate the features

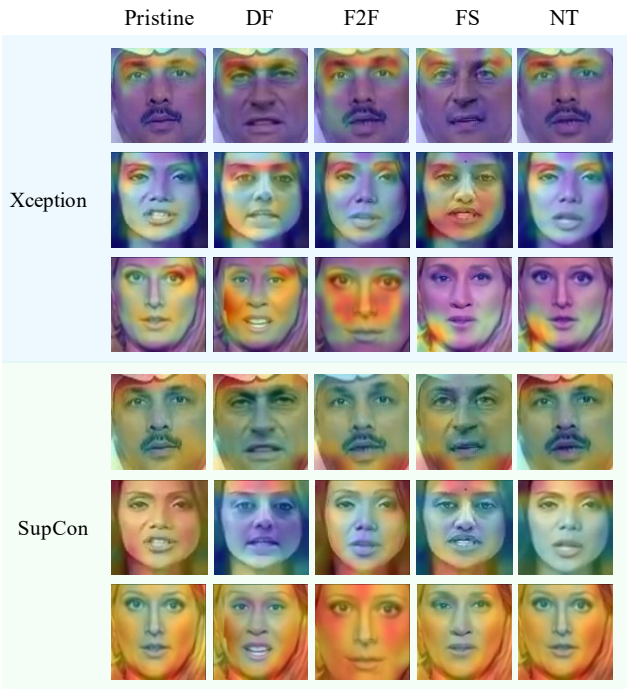

Figure 4: Heatmap for visualization of feature importance comparison from proposed Xception (top) and SupCon (bottom) mode.

from both models, which makes the features of the manipulated images to cluster in the middle, indicating the impact on classification. However, considering the feature spaces of different networks exhibiting different properties, it can be asserted that the feature level fusion needs detailed analysis. We thus resort to fusing the scores from the networks to benefit from the complementary nature of the two networks as discussed in the section below.

5. Fusion of SupCon and Xception

As different architectures learn different features, one can deduce the complementary nature of the features learnt and decisions made. Xception [15] has been reported to perform well for the DeepFakes detection task in the state of the art while our proposed approach learns different features. Motivated by such observations and backed by the explainability analysis provided in Section 4, we propose to fuse these two networks to obtain a better decision. We resort to using the decision scores fusion from each network due to the reasons discussed in the prior section. Furthermore, as the scores may be ideal for different tasks due to the nature of learning and to maximally use the scores from each network, we propose to employ a weighted fusion of the final score from the last activation layer. The architecture of our proposed fusion for DeepFakes detection is presented in Figure 6.

Considering the final scores $[s_0, s_1, s_2, s_3]$ from SupCon where s_0, s_1, s_2 s_3 is the prediction score for *pristine* (non-manipulated) and three known *DeepFakes*. Further, asserting the nature of cross-entropy loss, which provides high score to a particular class, we select the maximal score from the set of scores corresponding to manipulated images as a candidate [4] for fusion as below:

$$s_{real} = s_0$$
$$s_{fake} = \max(s_1, s_2, s_3)$$

Similarly, we obtain the final scores from Xception model as x_{real} and x_{fake}. Enforcing a linear combination of scores to obtain a final score f_{real} and f_{fake} for real and manipulated images respectively, we formulate the weighted combination using weights for proposed SupCon model as w_s and Xception model as w_x, we get our final fusion scores f_{real} and f_{fake}.

$$f_{real} + f_{fake} = 1$$
$$\begin{bmatrix} fs_{real} \\ fs_{fake} \end{bmatrix} = w_s \times \begin{bmatrix} s_{real} \\ s_{fake} \end{bmatrix} + w_x \times \begin{bmatrix} x_{real} \\ x_{fake} \end{bmatrix} \quad (3)$$

We adopt Greedy optimization of weights w_s and w_x to maximize the detection accuracy based on empirical trials as explained in the section below.

6. Experiments

To validate the proposed approach, we employ the FaceForensics++ consisting of 1000 original videos and corresponding number of manipulated videos consisting of 1000 videos for each of the subsets - DeepFakes (DF) [2], Face2Face (F2F) [63], FaceSwap (FS) [3] and NeuralTextures (NT) [62]. We choose c23 compression videos for our experiments balancing the size and quality of the videos [5]. We extract the frames from each of the videos resulting in 150000 total images, first 30 frames for each video.

As our experiments are focused on detecting the manipulated face region alone, we detect and crop the face region using Multi-task Cascade Convolutional Neural Networks (MTCNN) [69]. We allow loose cropping of the face region to capture the entire silhouette of the face region against the tight cropping. The detected faces were further resized to a standard size of 224×224 pixels for use in SupCon and 299×299 pixels for use in Xception network to match the input sizes of the Efficient-B0 and Xception networks.

6.1. Experimental Protocol

The total set of videos from each set of pristine and manipulated sets are separated into 600, 200, 200 for training, validation, and test in a disjoint and non-overlapping

[4] One can also use a linear combination of different scores.

[5] All the experiments were conducted on Python 3.6 environment with two GPUs on Ubuntu OS with one NVIDIA GRID V100D-8Q with 8GB of RAM and another combination of two NVIDIA TU102 with 16GB of RAM for training SupCon. We adopted the Pytorch framework [4] for developing Deep Learning models.

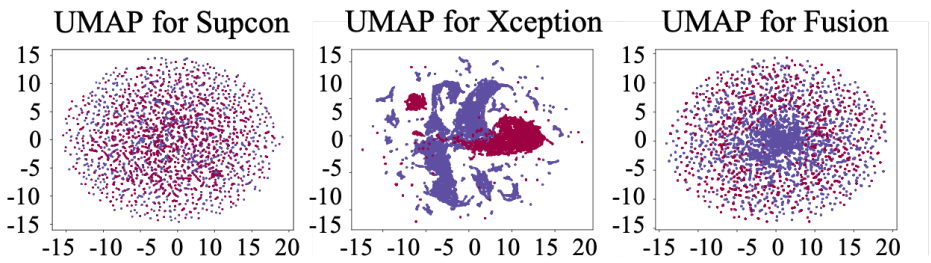

Figure 5: UMAPs of features for Supcon, Xception and fusion models. The purple dots correspond to manipulated images and the red ones correspond to pristine images.

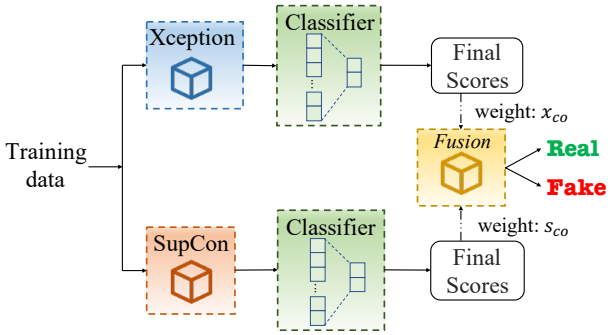

Figure 6: Framework for proposed approach using score level fusion of SupCon and Xception scores for DeepFakes detection.

manner. Considering the problem of open-set classification where the testing videos are unseen, we employ the pristine (original) videos and three manipulated videos in training for each case and retain one unused manipulated set as the unseen/unknown set. Thus, the results are reported on four different combinations: (i) Original + DeepFakes + Face2Face + FaceSwap *referred as (DF + F2F + FS)*, (ii) Original + DeepFakes + Face2Face + NeuralTextures *referred as (DF + F2F + NT)*, (iii) Original + DeepFakes + FaceSwap + NeuralTextures *referred as (DF + FS + NT)*, (iv) Original + Face2Face + FaceSwap + NeuralTextures *(F2F + FS + NT)*. We, therefore, report the results as Closed-Set-Classification (CSC) Accuracy and True-Open-Set-Classification (TOSC) Accuracy where the first metric is an accuracy metric when the test class is seen during the training, and the second metric corresponds to a pure open-set case when the test class is unseen at training. Furthermore, we also report the Open-Set-Classification (OSC) accuracy taking into account the combined accuracy when the unknown data is tested with known train classes.

- **Closed-Set-Classification (CSC) Accuracy:** For a set of n known DeepFakes classes in the training set, say $df_1, df_2, \ldots df_n$ and real videos r, let TP_i, TN_i, FP_i, and FN_i respectively denote the true positive, true negative, false positive, and false negative for the i-th Closed Set Class, where $i \in \{1, 2, ..., C\}$ and $C \in r, df_1, df_2, \ldots df_n$

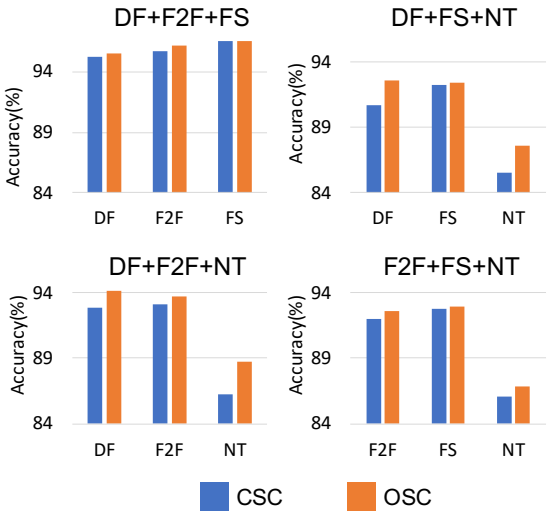

Figure 7: CSC and OSC accuracy obtained from proposed SupCon model.

denotes the number of known classes in training and testing, we can obtain the following CSC as:

$$CSC = \frac{\sum_{i=1}^{C}(TP_i + TN_i)}{\sum_{i=1}^{C}(TP_i + TN_i + FP_i + FN_i)}. \quad (4)$$

- **True-Open-Set-Classification (TOSC) Accuracy**: A trivial modification of the above metric in the unknown test setting can be obtained by simply setting the open set class to known training classes. Let TP_i, TN_i, FP_i, and FN_i respectively denote the true positive, true negative, false positive, and false negative for the i-th Open Set Class, where $i \in \{1, 2, ..., C, C+1\}$ and C denotes the number of known classes in training and the test class is represented by $C+1$. As long as the test class $C+1$ is classified as a manipulated class when the test class is manipulated and non-manipulated class otherwise, the TOSC can be presented as:

$$TOSC = \frac{\sum_{i=1}^{C}(TP_i + TN_i)}{\sum_{i=1}^{C}(TP_i + TN_i + FP_i + FN_i)} \quad (5)$$
$$where \iff C+1 \subset \{1, 2, ..., C\}.$$

Table 1: Classification results (CSC and OSC) obtained from the proposed SupCon model trained on three known attack classes in the training set and one unknown attack class in the testing set. *Indicates TOSC where the test class is not employed in the training set.

	Train Classes							
Test	F2F + FS + NT		DF + FS + NT		DF + F2F + NT		DF + F2F + FS	
Class	**CSC**	**OSC**	**CSC**	**OSC**	**CSC**	**OSC**	**CSC**	**OSC**
DF	-	78.74*	90.70	92.59	92.86	94.14	95.24	95.54
F2F	92.03	92.58	-	58.82*	93.11	93.70	95.72	96.14
FS	92.81	92.91	92.26	92.36	-	47.55*	96.52	96.57
NT	86.09	86.86	85.51	87.55	86.24	88.76	-	54.61*

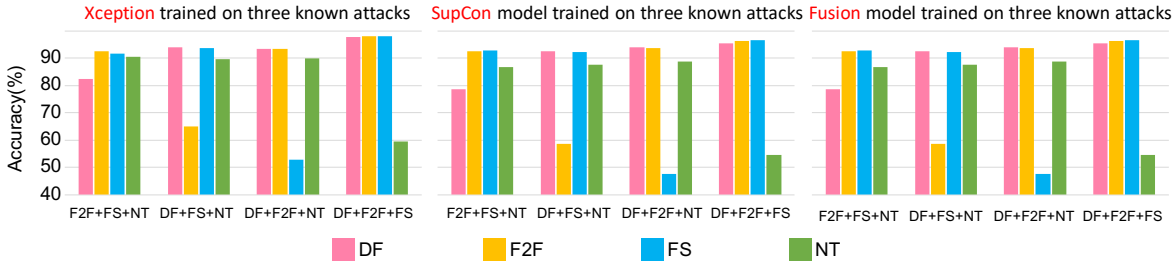

Figure 8: OSC of Xception, SupCon and Fusion model trained with three known attacks.

- **Open-Set-Classification (OSC) Accuracy:** If TU and FU respectively denote the correct and false reject for unknown classes, OSC can be defined as:

$$OSC = \frac{\sum_{i=1}^{C}(TP_i + TN_i) + TU}{\sum_{i=1}^{C}(TP_i + TN_i + FP_i + FN_i) + (TU + FU)}. \quad (6)$$

6.2. Results: Proposed SupCon Model

We first present the baseline results of closed-set classification in Table 1, along with the open set classification accuracy to empirically validate the proposed approach of SupCon for DeepFakes detection. As noted from Table 1, the proposed approach performs with more than 93.47% average accuracy on all closed datasets with an exception to NT, where the accuracy results in around 85.94% on average. The same results are also presented in Figure 7 for the ease of reading. While the results are not surprising, given that the supervised contrastive learning has shown its potential in other applications [32], the obtained results in closed set experiments validate the potential of supervised contrastive learning for detecting DeepFakes. Furthermore, as noted from the open set classification results in Table 1, the accuracy in detecting DF as an unknown class in 78.74% in the true unknown class setting (TOSC), while the same accuracy drops to 58.82%, 47.55% and 54.61% for F2F, FS and NT respectively. However, the combined accuracy (OSC) when the unknown test class is combined with the known training class improves consistently, indicating no deterioration of the proposed approach for known test classes. Such a scenario is expected in real-life cases where the performance of the known classes need to be maintained while detecting unknown classes and the proposed approach fits the criteria in few potential applications.

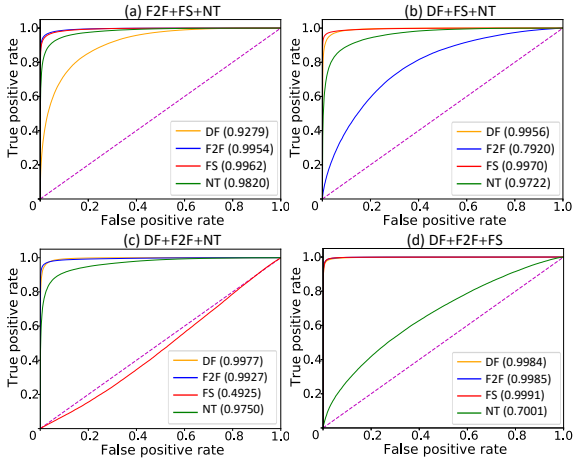

Figure 9: ROC curve for fusion model trained on various combinations (noted on the top) of training classes. AUCs noted in the brackets in legend.

We further validate the proposed approach against Xception provided in Table 2, and the same is illustrated in Figure 8 for providing the reader with a state-of-the-art comparison [6]. As noted from Table 2, we note the accuracy of proposed approach is comparable against the Xception model.We exit the training of the proposed approach after 30 epochs and we hypothesize that more epochs would result in performance gain due to learning better contrasts. This aspect, however, shall be studied in future works.

[6]The detailed evaluation of other state-of-the-art approaches are provided in supplementary materials on this link due to page constraints

Table 2: Test accuracy (in %) of proposed SupCon model, Xception and proposed fusion model. Highlighted rows indicate TOSC accuracy where the reported test class is not in training set and blue arrows denote increase in accuracy as compared to either of the models trained independently.

Training Class	Test Class	SupCon (Proposed)	Xception	Fusion (Proposed)
F2F + FS + NT	DF	78.74	82.73	83.99 ↑
	F2F	92.58	93.21	94.36 ↑
	FS	92.91	93.31	94.74 ↑
	NT	86.86	91.88	92.45 ↑
DF + FS + NT	DF	92.59	94.71	95.44 ↑
	F2F	58.82	64.69	64.69
	FS	92.36	94.71	95.39 ↑
	NT	87.55	90.17	91.82 ↑
DF + F2F + NT	DF	94.14	94.26	95.31 ↑
	F2F	93.70	93.49	94.69 ↑
	FS	47.55	49.74	49.77 ↑
	NT	88.76	90.88	91.82 ↑
DF + F2F + FS	DF	95.54	97.79	97.95 ↑
	F2F	96.14	97.91	98.04 ↑
	FS	96.57	98.20	98.29 ↑
	NT	54.61	55.59	55.59

6.3. Results: Proposed Fusion

Table 2 presents the results obtained from the proposed fusion approach in comparison to the proposed SupCon model and Xception model. As it can be noted, our fusion model consistently performs better than a single model, while reaching at-least the lower bound of the performance of Xception model in three cases (not accompanying ↑). The obtained results with a consistent gain indicate the complementarity of our proposed approach in improving the generalizability. We further present the Receiver Operating Characteristic (ROC) curves of the proposed fusion model to indicate the overall performance in Figure 9.

6.4. Comparison with State-of-the-art Methods

We further benchmark our proposed approaches against four other state-of-the-art works which address the open-set problem in DeepFakes detection to provide a fair benchmark. Specifically, we benchmark our approach against a convolutional LSTM based residual network - CLRNet [60], Transfer learning-based Autoencoder with Residuals (TAR) [35], Generalized Zero and Few-Shot Transfer approach [7] and Xception network [15]. As noted from the Table 3, the proposed approach provides either better performance or, in some cases, provides comparable performance compared to existing works. While a higher accuracy is noted from Lee *et al*. [35], they adopt a different protocol by using one known class for training and testing one unknown class. While our proposed method is promising as they have a consistent performance with other state-of-the-art methods, our approach performs better than existing methods. For instance, FS has the lowest accuracy results on all of these models, while DF obtains the highest detection accuracy. The lower performance of FS is worse than the random guess, and this needs to be further investigated in future works.

Table 3: TOSC Accuracy obtained on the unknown classes from various state-of-the-art approaches against proposed approaches.

Model	DF	F2F	FS	NT
CLRNet [60]	50.12	53.73	50.00	**69.75**
TAR [35] (Lowest /Highest)	75.25 (50.8/ 99.70)†	**72.90** (50.0/ 75.30)†	**51.65** (50.1/ 52.20)†	-
DDT [7]	78.82	-	-	64.10
Xception [15]*	82.73	64.69	49.74	55.59
SupCon(Ours)	78.74	58.82	47.55	54.61
Fusion(Ours)	**83.99**	64.69	49.77	55.59

* Our implementation of the method. † indicates the results not directly comparable to our approach as the best accuracy is reported on one-known-training-class and NT was not employed in any training.

7. Limitations of Current Work

While simple score level fusion provides promising results in our work, it is necessary to study the feature level fusion and feature selection approach in future works to mitigate the shortcomings in generalization performance. Specifically, the proposed approaches, along with the Xception model fail to generalize well in three cases of true open set protocols (F2F, FS, and NT, with the performance of FS being the lowest). This can be, to a certain degree, attributed to the fact that the FS works on swapping the face region including the silhouette (as shown in Fig 4). Employing the complete scene information, the background, or looking at temporal differences in the video frames can further improve the performance and this has to be studied in future works.

8. Conclusion

There is an imperative need for a generalized DeepFakes detection method to deal with the newer manipulation methods in visual media. We have presented a DeepFakes detection method based on supervised contrastive learning to provide a generalizable and explainable model. Using the explainability as the basis, we have further proposed a fusion model using Xception architecture. Experiments conducted on publicly available FaceForensics++ dataset demonstrate the potential of our method and show comparable performance against state-of-the-art generalized DeepFakes detection algorithms. Future works on developing the idea need to investigate the cross-database performance for DeepFakes detection, analyse different architectures for the SupCon model, and investigate the feature level fusion of the complementary networks.

References

[1] Deepfacelab. https://github.com/iperov/DeepFaceLab. 2021-10-25.

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

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

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

[5] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. The state of deepfakes: Landscape, threats, and impact. *Amsterdam: Deeptrace*, 2019.

[6] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv*, pages 1–17, 2020.

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

[8] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. *arXiv*, pages 1–6, 2020.

[9] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5012–5019. IEEE, 2021.

[10] Lucy Chai, David Bau, Ser Nam Lim, and Phillip Isola. What Makes Fake Images Detectable? Understanding Properties that Generalize. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12371 LNCS:103–120, 2020.

[11] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil M. Robertson, and Stan Z. Li. Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection. *IEEE Transactions on Information Forensics and Security*, 15(2):578–593, 2020.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv*, (Figure 1), 2020.

[13] Zehao Chen and Hua Yang. Attentive Semantic Exploring for Manipulated Face Detection. 1:1–5, 2020.

[14] Zehao Chen and Hua Yang. Manipulated face detector: Joint spatial and frequency domain attention network. *arXiv preprint arXiv:2005.02958*, 2020.

[15] François Chollet. Xception: Deep learning with depthwise separable convolutions. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:1800–1807, 2017.

[16] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *arXiv*, X(X):1–17, 2019.

[17] V Conotter, E Bodnari, and G Boato. PHYSIOLOGICALLY-BASED DETECTION OF COMPUTER GENERATED FACES IN VIDEO Department of Information Engineering and Computer Science University of Trento , Trento (ITALY) Dartmouth College Department of Computer Science Hanover NH 03755 (USA). *International Conference on Image Processing(ICIP)*, pages 248–252, 2014.

[18] 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.

[19] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[20] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Randaugment Le. Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019.

[21] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the Detection of Digital Face Manipulation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5780–5789, 2020.

[22] Xinyi Ding, Zohreh Raziei, Eric C. Larson, Eli V. Olinick, Paul Krueger, and Michael Hahsler. Swapped face detection using deep learning and subjective assessment. *arXiv*, 8, 2019.

[23] Nhu-Tai Do, In-Seop Na, and Soo-Hyung Kim. Forensics face detection from gans using convolutional neural network. In *2018 International Symposium on Information Technology Convergence (ISITC 2018), South Korea*, 2018.

[24] Ricard Durall, Margret Keuper, Franz Josef Pfreundt, and Janis Keuper. Unmasking DeepFakes with simple features. *arXiv*, 2019.

[25] 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. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pages 1721–1729, 2019.

[26] Sheldon Fung, Xuequan Lu, Chao Zhang, and Chang-Tsun Li. Deepfakeucl: Deepfake detection via unsupervised contrastive learning. *arXiv preprint arXiv:2104.11507*, 2021.

[27] Ipek Ganiyusufoglu, L Minh Ngô, Nedko Savov, Sezer Karaoglu, and Theo Gevers. Spatio-temporal features for generalized detection of deepfake videos. *arXiv preprint arXiv:2010.11844*, 2020.

[28] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[29] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. DeepFake Detection by Analyzing Convolutional Traces. *arXiv*, 2020.

[30] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Fighting deepfake by exposing the convolutional traces on images. *IEEE Access*, 8:165085–165098, 2020.

[31] David Guera and Edward J. Delp. Deepfake Video Detection Using Recurrent Neural Networks. *Proceedings of*

AVSS 2018 - 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2019.

[32] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Dilip Krishnan, and Ce Liu. Supervised contrastive learning. *arXiv*, (NeurIPS):1–23, 2020.

[33] 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.

[34] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[35] Sangyup Lee, Shahroz Tariq, Junyaup Kim, and Simon S Woo. Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 351–366. Springer, 2021.

[36] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5000–5009, 2020.

[37] Yuezun Li, Ming Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv*, 2018.

[38] Yuezun Li and Siwei Lyu. Exposing DeepFake Videos by Detecting FaceWarping Artifacts. *arXiv*, 2018.

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 772–781, 2021.

[40] Zhengzhe Liu, Xiaojuan Qi, and Philip H.S. Torr. Global Texture Enhancement for Fake Face Detection in the Wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8057–8066, 2020.

[41] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[42] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of GAN-Generated Fake Images over Social Networks. *Proceedings - IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018*, pages 384–389, 2018.

[43] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs Leave Artificial Fingerprints? *Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019*, pages 506–511, 2019.

[44] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision Workshops, WACVW 2019*, pages 83–92, 2019.

[45] Scott McCloskey and Michael Albright. Detecting GAN-Generated Imagery Using Saturation Cues. *Proceedings - International Conference on Image Processing, ICIP*, 2019-Septe:4584–4588, 2019.

[46] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[47] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: A deepfake detection method using audio-visual affective cues. *arXiv preprint arXiv:2003.06711*, 2020.

[48] 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.

[49] Timo Ojala, Matti Pietikainen, and David Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th international conference on pattern recognition*, volume 1, pages 582–585. IEEE, 1994.

[50] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4318–4327, 2020.

[51] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12357 LNCS:86–103, 2020.

[52] Andrea Macarulla Rodriguez, Marissa Koopman, Andrea Macarulla Rodriguez, and Zeno Geradts. Detection of Deepfake Video Manipulation. *Imvip*, (December):133–136, 2018.

[53] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. *arXiv*, 2019.

[54] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *arXiv*, 2019.

[55] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[56] Kritaphat Songsriin and Stefanos Zafeiriou. Complement face forensic detection and localization with facial landmarks. *arXiv*, 2019.

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

[58] Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6):1635–1650, 2010.

[59] Shahroz Tariq, Sangyup Lee, and Simon Woo. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of the Web Conference 2021*, pages 3625–3637, 2021.

[60] Shahroz Tariq, Sangyup Lee, and Simon S Woo. A convolutional LSTM based residual network for deepfake video detection. *arXiv preprint arXiv:2009.07480*, 2020.

[61] The Washington Post. How misinformation helped spark an attempted coup in gabon. https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon, 2020. Accessed: 2020-02-13.

[62] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv*, 2019.

[63] 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.

[64] VICE. We've just seen the first use of deepfakes in an indian election campaign. https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp, 2020. Accessed: 2020-02-18.

[65] Sheng Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot.. For Now. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8692–8701, 2020.

[66] Xi Wu; Zhen Xie; YuTao Gao; Yu Xiao. SSTNET : Detecing manipulated faces through spatial, stegenalysis and temporal features YuTao Gao Yu Xiao Alibaba Group China. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2952–2956, 2020.

[67] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[68] Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. *arXiv*, pages 7556–7566, 2018.

[69] 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.

[70] Xu Zhang, Svebor Karaman, and Shih Fu Chang. Detecting and simulating artifacts in GAN fake images. *arXiv*, 2019.

[71] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-Stream Neural Networks for Tampered Face Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017-July:1831–1839, 2017.