Chapter 10
3D CNN Architectures and Attention Mechanisms for Deepfake Detection

Ritaban Roy, Indu Joshi, Abhijit Das, and Antitza Dantcheva

Abstract Manipulated images and videos have become increasingly realistic due to the tremendous progress of deep convolutional neural networks (CNNs). While technically intriguing, such progress raises a number of social concerns related to the advent and spread of fake information and fake news. Such concerns necessitate the introduction of robust and reliable methods for fake image and video detection. Toward this in this work, we study the ability of state-of-the-art video CNNs including 3D ResNet, 3D ResNeXt, and I3D in detecting manipulated videos. In addition, and toward a more robust detection, we investigate the effectiveness of attention mechanisms in this context. Such mechanisms are introduced in CNN architectures in order to ensure that robust features are being learnt. We test two attention mechanisms, namely SE-block and Non-local networks. We present related experimental results on videos tampered by four manipulation techniques, as included in the FaceForensics++ dataset. We investigate three scenarios, where the networks are trained to detect (a) all manipulated videos, (b) each manipulation technique individually, as well as (c) the veracity of videos pertaining to manipulation techniques not included in the train set.

R. Roy
BITS, Pilani, India
e-mail: f2015842@pilani.bits-pilani.ac.in

I. Joshi
IIT Delhi, Delhi, India
e-mail: indu.joshi@cse.iitd.ac.in

A. Das (✉)
Thapar University, Patiala, India
e-mail: abhijit.das@thapar.edu

A. Dantcheva
Inria Sophia Antipolis, Valbonne, France
e-mail: antitza.dantcheva@inria.fr

C. Rathgeb et al. (eds.), *Handbook of Digital Face Manipulation and Detection*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-87664-7_10

10.1 Introduction

Manipulated images date back to the creation of the first photograph in the year 1825 [18]. Related manipulation techniques have been widely driven by profit stemming from identity theft, age deception, illegal immigration, organized crime, and espionage, inflicting negative consequences on businesses, individuals, and political entities. While forgery was associated with a slow, painstaking process usually reserved for experts, we are entering new levels of manipulation of images and video, where deep learning and related *manipulation* are streamlined to reduce costs, time, and skill needed to doctor images and videos. Automated generation and manipulation of audio, image and video bares highly exciting perspectives for science, art and video productions, e.g., video animation, special effects, reliving already passed actors.

While highly intriguing from computer vision perspective, *deepfakes* entail a number of challenges and threats, given that (a) such manipulations can fabricate animations of subjects involved in actions that have not taken place and (b) such manipulated data can be circumvented nowadays rapidly via social media. Particularly, we cannot trust anymore, what we see or hear on video, as deepfakes betray sight and sound, the two predominantly trusted human innate senses [44]. Given that (i) our society relies heavily on the ability to produce and exchange legitimate and trustworthy documents, (ii) sound and images have recorded our history, as well as informed and shaped our perception of reality, e.g., axioms and truths such as "I'll believe it when I see it." "Out of sight, out of mind." "A picture is worth a thousand words". (iii) Social media has catapulted online videos as a mainstream source of information; deepfakes pose a threat of distorting what is perceived as reality. To further fuel concern, deepfake techniques have become open to the public via phone applications such as FaceApp[1], ZAO[2] and Wombo[3]. Further, *digital identity*[4], associated to the entire collection of information generated by a person's online activity including usernames and passwords, photographs, online search activities, birth date, social security becomes highly vulnerable, with deepfakes entailing the premise to inflict severe damage. Additional social threats [12, 17] can affect domains such as journalism, education, individual rights, democratic systems and have intrigued a set of journalists[5, 6, 7, 8].

[1] https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341.

[2] https://apps.apple.come/cn/app/id146519927.

[3] https://www.wombo.ai/.

[4] https://www.indrastra.com/2018/01/Digital-Identity-Gateway-to-All-Other-Use-Cases-004-01-2018-0034.html.

[5] https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/.

[6] https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html.

[7] https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin.

[8] https://www.cnbc.com/2019/10/14/what-is-deepfake-and-how-it-might-be-dangerous.html.

We differentiate two cases of concern: the first one has to do with *deepfakes being perceived as real*, and the second relates to *real videos being misdetected for fake*, the latter referred to as "liar's dividend". Given such considerations, e.g., video evidence becomes highly questionable.

Recent research on deepfake generation proposed approaches, where forged videos are created based on a *short video* of the source person [30, 48], as well as from a *single ID photo* [5] of the source person. In addition, fully synthesized *audio-video* images are able to replicate synchronous speech and lip movement [46] of a target person. Hence deepfakes coerce the target person in a video to reenact the dynamics of the source person.

Two deepfake-schemes have evolved, corresponding to *head puppetry* (the dynamics of a head from a source person are synthesized in a target person), as well as face swapping (the whole face of a target person is swapped with that of a source person). Lip syncing (the lip region of the target person is reenacted by the lip region of a source person) falls in the first category. Currently such manipulations include subtle imperfections that can be detected by humans and, if trained well, by computer vision algorithms [3, 32, 33]. Toward thwarting such attacks, early multimedia forensics based detection strategies have been proposed [3, 4, 16, 41]. Such strategies, although essential, cannot provide a comprehensive solution against manipulated audio, images, and video. Specifically, the detection of deepfakes is challenging for several reasons: (a) it evolves a "cat-and-mouse-game" between the adversary and the system designer, (b) deep models are highly domain-specific and likely yield big performance degradation in cross-domain deployments, especially with large train-test domain gap.

The *manipulation scenario* of interest in this work has to do with a face video or expressions of a *target person* being superimposed to a video of a *source* person, widely accepted and referred to as *deepfake*.

Contributions

Motivated by the above, this work makes following contributions.

(i) We compare state-of-the-art *video* based techniques in detecting deepfakes. Our intuition is that current state-of-the-art forgery detection techniques [1, 8, 14, 19, 39, 40] omit a pertinent clue, namely, *motion*, by investigating only spatial information. It is known that generative models have exhibited difficulties in preserving appearance throughout generated videos, as well as motion consistency [42, 51, 54, 57]. Hence, we here show that using 3D CNNs indeed outperforms state of the art image-based techniques.

(ii) We show that such models trained on known manipulation techniques generalize poorly to tampering methods outside of the training set. Toward this, we provide an evaluation, where train and test sets do not intersect with respect to manipulation techniques.

(iii) We determine the efficacy of two attention mechanisms, namely SE-block and Non-local networks by comparing the number of parameters, inference time, and classification performance for deepfake detection. We find that a non-local

neural network indeed improves the classification accuracy of 3D CNNs without introducing significant computational overhead.

(iv) Lastly, we analyze the correlation matrix of learnt features, as well as activations of Seg-Grad-Cam [53] to provide insight on how attention mechanisms work.

We note that this chapter extends the work of Wang and Dantcheva [60] by contributions (iii) and (iv).

10.2 Related Work

A very recent survey has revisited image and video manipulation approaches and early detection efforts [49]. An additional comprehensive survey paper [63] reviews manipulations of images, graphs, and text.

Generative adversarial networks (GANs) [20] have enabled a set of face manipulations including identity [28, 35], facial attributes [61], as well as facial expressions [27, 34, 57–59].

10.2.1 Deepfake Detection

While a number of *manipulation-detection-approaches* are *image-based* [1, 40], others are targeted toward *video* [3, 33, 41] or jointly toward audio and video [31]. We note that although some **video-based approaches** might perform better than image-based ones, such approaches are only applicable to particular kinds of attacks. For example, many of them [3, 33] may fail, if the quality of the eye area is not sufficiently good or the synchronization between video and audio is not sufficiently natural [32].

Image-based approaches are general-purpose detectors, for instance, the algorithm proposed by Fridrich and Kodovsky [19] is applicable to both steganalysis and facial reenactment video detection. Rahmouni et al. [39] presented an algorithm to detect computer-generated images, which was later extended to detecting computer-manipulated images. However, performance of such approaches on new tasks is limited compared to that of task-specific algorithms [40].

Agarwal et al. exploited both facial identity as well as behavioral biometrics information provided by the temporal component of videos to classify a video as real or fake [2]. Cozzolino et al. used temporal facial features to learn behavior of a person and use this as an identifier to compare characteristics in the presented video and verify the claim of identity [15]. Guarnera et al. argued that deepfake videos contain a forensic trait pertaining to the generative model used to create them. Specifically, they showed that convolutional traces are instrumental in detecting deepfakes [22]. Khalid and Woo [29] posed deepfake detection as an anomaly detection problem and used variational auto-encoder for detecting deepfakes. Hernandez-Ortega [24] proposed a deepfake detection framework based on physiological measurement, namely, heart

rate using remote photoplethysmography (rPPG). Trinh et al. [50] utilized dynamic representations (i.e., prototypes) to explain deepfake temporal artifacts. Sun et al. [45] attempted to generalize forgery face detection by proposing a framework based on meta-learning. Tolosana et al. [49] revisited first and second DeepFake generations w.r.t. facial regions and fake detection performance.

We show in this work that such algorithms are indeed challenged, if confronted with manipulation techniques outside of the training data.

Rössler et al. [40] presented a comparison of existing handcrafted, as well as deep neural networks (DNNs), which analyzed the **FaceForensics++** dataset and proceeded to detect adversarial examples in an *image-based* manner. This was done for (i) raw data, (ii) high quality videos compressed by a constant rate quantization parameter equal to 23 (denoted as HQ), as well as (iii) low quality videos compressed by a quantization rate of 40 (denoted as LQ). There were two training settings used: (a) training on all manipulation methods concurrently, (b) individual training on each manipulation method separately. These two settings refer to the first two scenarios of interest in this work.

We summarize for training setting (a), which is the more challenging setting (as indicated by lower related detection rates).

1. **Raw data**: It is interesting to note that the correct detection rates for all seven compared algorithms ranged between 97.03 and 99.26%. The highest score was obtained by the XceptionNet [13].
2. **HQ**: High quality compressed data was detected with rates, ranging between 70.97 and 95.73% (XceptionNet).
3. **LQ**: Intuitively low quality compressed data had the lowest detection rates with 55.98–81% (XceptionNet).

We here focus on the LQ-compression as the most challenging setting.

We note that reported detection rates pertained to the analysis of a facial area with the dimension 1.3 times the cropped face. Analyzing the full frame obtained lower accuracy.

A challenge, not being addressed by Rössler et al. has to do with the generalization of such methods. When detection methods, as the presented ones are confronted with adversarial attacks, outside of the training set, such networks are challenged. This has to do with the third scenario of interest in this chapter.

10.2.2 Attention Mechanisms

Attention mechanisms are designed to identify and focus on salient information, which can facilitate improved decisions. Deepfake videos are acquired in uncontrolled conditions and can include a number of artificially created objects in the background (e.g., news-banners). We hypothesize that attention mechanisms are instrumental in facilitating improved classification accuracy of a deepfake detector

by enabling the model to focus on discriminative information. Additionally, visualization of attention maps is beneficial in interpretation of the taken decision.

The understanding about attention can be derived from Nadaraya-Watson's regression model [37, 62]. Given the paired training data $\{(x_1, y_1), (x_2, y_2)...(x_n, y_n)\}$, for a given test example x, a regression model predicts the target value $\hat{y}$ as

$$\hat{y} = \sum_{k=1}^{n} \alpha(x, x_k) y_k \tag{10.1}$$

i.e., the target value is a weighted average of training instances. Here, the weight $\alpha(x, x_k)$ signifies the relevance of training instance x_k for making a prediction for x. Attention mechanisms in deep models are analogous to Nadaraya-Watson's regression model, as such models are similarly designed to learn a weighting function.

Attention models incorporate an encoder-decoder architecture, solving the pitfall of auto-encoder by allowing the decoder to access the entire encoded input sequence. Attention aims at automatically learning an attention weight, which captures the relevance between the *encoder hidden state*, i.e., candidate state and the *decoder hidden state* i.e., the query state. The seminal work on attention was proposed by Bahdanau et al. [6] for a sequence-to-sequence modeling task. Attention modeling has evolved to different types of attention based on the category of input and output, as well as application domain. While the input of an attention model constitute an image, sequence, graph, or tabular data and the output is represented by an image, sequence, embedding, or a scalar. We note that attention can be categorized based on the number of sequences, number of abstraction levels, number of positions, as well as number of representations [11]. We proceed to explain such types in detail.

With respect to **number of sequences**, attention can be of three types, namely, distinctive, co-attention, and self attention. While in distinctive attention candidate and query states belong to two distinct input and output sequences, in self attention [38, 52] the candidate and query states belong to the same sequence. In contrast, co-attention accepts multiple input sequences as input at the same time and jointly produces an output sequence.

Considering **number of abstraction**, attention can be divided into two types of levels, namely, single-level and multilevel. In single-level attention weights are computed only for the original input sequence, whereas in multilevel there are lower and higher level of abstraction, works can be organized in top-down or bottom-up approaches.

While considering the **number of positions**, attention can be of two types, soft/global and hard/local. Hard attention requires the weights to be binary; for instance, a model that crops the image toward naturally discarding non-necessary details [21]. A major limitation of hard attention is that it is implemented using stochastic non-differentiable algorithms [7, 36]. As a result, models employing it cannot be trained in an end-to-end manner. Deviating from this, models employing soft attention take an image or video as input and soft-weigh the region of interest [26, 55]. Soft weighing is ensured by employing either sigmoid or softmax after the

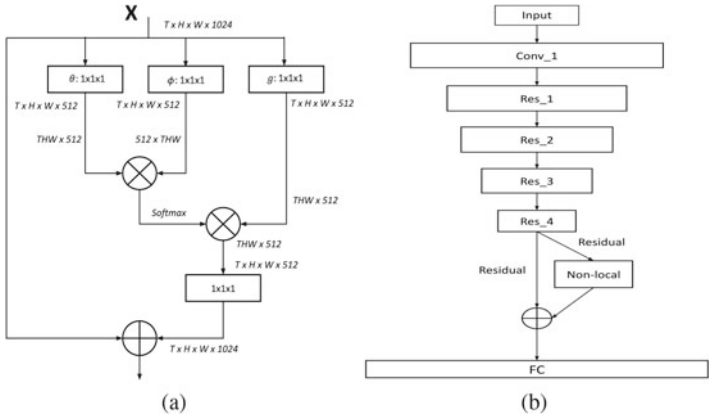


Fig. 10.1 Schematic diagrams of **a** non-local block, **b** non-local block in the backbone architecture

attention gates. This allows weights to be real valued and the objective function to be differentiable.

Based on **number of representations** we have multi-representational and multidimensional attention. While in the former different aspects of the input are considered, in the latter focus is placed on determining the relevance of each dimension of the input.

Finally, with respect to the type of **architecture**, related attention models can be implemented as encoder-decoder, transformer, and memory networks. An encoder-decoder based attention model takes any input representation and reduces it to a single fixed length, a transformer network aims to capture global dependencies between input and output, and in memory networks facts that are more relevant to the query are filtered out.

Application domains of attention include (i) natural language processing, (ii) computer vision, (iii) multi-modal tasks, (iv) graphical systems, and (v) recommender systems. Visual attention brings to the fore a vector of importance weights; in order to predict or infer one element, e.g., a pixel in an image, we estimate using the attention vector how meaningful it is. In particular in this scenario, attention modules are designed to indicate decisive regions of an input, for the task in hand. The output of an attention module is a vector, representing relative importance. This vector is then used to re-weight network parameters, so that pertinent characteristics have higher weights. Consequently, an attention module boosts the model's performance in a targeted task. For this work we introduce a self attention, soft attention, single-level, multidimensional attention for deepfake detection.

We proceed to describe two promising modules used extensively and successfully in image and video processing applications, and which we employ in this chapter, viz., *non-local block*, which is based on transformer network and *squeeze and excitation* that is based on an encoder-decoder network.

Table 10.1 Architecture of 3D ResNet-101 with non-local block

Layer name	Output size	Architecture
Conv_1	$16 \times 112 \times 112$	$7 \times 7 \times 7$, 3, stride (1, 2, 2)
Res_1	$16 \times 56 \times 56$	$3 \times 3 \times 3$ maxpool, stride 2 $\begin{pmatrix} 1 \times 1 \times 1, 64 \\ 3 \times 3 \times 3, 64 \\ 1 \times 1 \times 1, 256 \end{pmatrix} \times 3$
Res_2	$8 \times 28 \times 28$	$\begin{pmatrix} 1 \times 1 \times 1, 128 \\ 3 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 512 \end{pmatrix} \times 4$
Res_3	$4 \times 14 \times 14$	$\begin{pmatrix} 1 \times 1 \times 1, 256 \\ 3 \times 3 \times 43, 256 \\ 1 \times 1 \times 1, 1024 \end{pmatrix} \times 23$
Res_4	$2 \times 7 \times 7$	$\begin{pmatrix} 1 \times 1 \times 1, 256 \\ 3 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 1024 \end{pmatrix} \times 3$
Non-local block	$2 \times 7 \times 7$	Fig. 10.1a
Avg pool & FC	$1 \times 1 \times 1$	Average pool and sigmoid

Non-local Block

The architecture of a non-local block [56] is based on the observation that convolutional and recurrent operations process only a local neighborhood. Consequently, these fail to capture long-range dependencies. To overcome this limitation of CNNs, non-local block performs a non-local operation to compute feature responses (see Fig. 10.1 and Table 10.1). A non-local operation is characterized by computing the response at a position as a weighted sum of features at all positions in the input feature maps.

Given that video processing requires access to information in distant pixels in space an time, computation of long-range dependencies is necessitated. Non-local operations enable a CNN to capture long-range dependencies and thus are highly beneficial in video processing. Formally, in the context of CNNs, a non-local operation is defined as

$$o_i = \frac{1}{C(x)} \sum_{\forall j} p(x_i, x_j) \;\; r(x_j), \tag{10.2}$$

where x and o denote the input and output feature, respectively. p represents a pairwise function that computes a relationship (e.g., affinity) between pixels i and j. r signifies a unary function, which computes a representation of input feature at pixel j. $C(x)$ is a normalization factor and is set as $C(x) = \sum_{\forall j} p(x_i, x_j)$.

In this chapter, the default choices of p and r are used. g is a linear embedding and is defined as $g(x) = W_g \; x_j$. Pairwise function is defined as

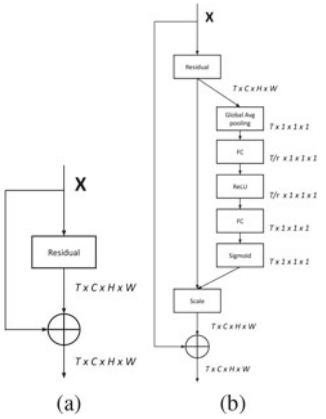


Fig. 10.2 **a** Residual block, **b** residual block after adding SE-block

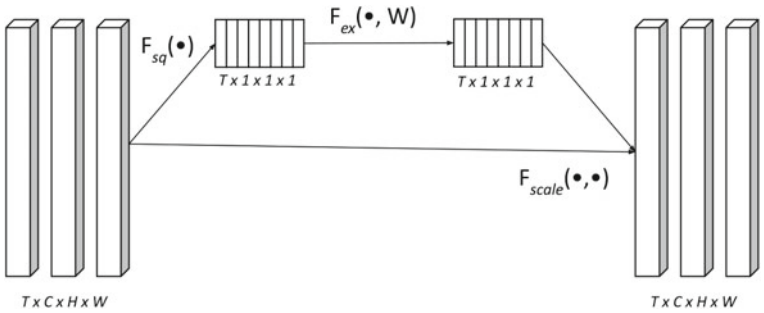


Fig. 10.3 Schematic diagram of SE-block showcasing the squeeze and excitation function

$$p(x_i, x_j) = e^{\alpha(x_i)^T \beta(x_j)}, \tag{10.3}$$

where $\alpha(x_i) = W_\alpha x_i$ and $\beta(x_j) = W_\beta x_j$ are the associated embeddings. This pairwise function is called embedded Gaussian and primarily computes dot-product similarity in the embedding space.

Squeeze and Excitation Block

The Squeeze and Excitation (SE) block [25] boosts the representational power of a CNN by modeling inter-dependencies between channels of the features learnt by it (see Fig. 10.2). As illustrated in Fig. 10.3, the SE-block comprises two operators: squeeze and excitation. While the *squeeze* operation aggregates features across spatial dimensions and creates a global distribution of channel-level feature response, the *excitation* operation is a self-gating mechanism that generates a vector of per-channel re-calibration weights. We proceed to define both operations.

Squeeze Operation. Let us assume that the input feature $X \in R^{W \times H \times C}$ is represented as $X = [x_1, x_2, \ldots x_C]$, where $x_i \in R^{W \times H}$. The squeeze operation exploits

(a) (b) (c) (d) (e)

Fig. 10.4 Sample frames from the FaceForensics++ dataset. From left to right: original source (large) and target (small) images, deepfakes, face2face, faceswap, neuraltextures

global spatial information by squeezing X through global average pooling and creating a channel descriptor, $z \in R^C$ where ith element of z is calculated as

$$z_i = F_{sq}(x_i) = \frac{1}{H \times W} \sum_{j=1}^{W} \sum_{k=1}^{H} x_i(j,k). \tag{10.4}$$

Excitation Operation. Exploits information acquired through squeeze operation to model dependency among channels through gating with sigmoid activation. Formally, squeeze operation is defined the following.

$$a = F_{ex}(z, w_1, w_2) = \sigma(w_2 \delta(w_1 z)), \tag{10.5}$$

where $w_1 \in R^{\frac{C}{r} \times C}$, $w_2 \in R^{C \times \frac{C}{r}}$. In this context a denotes the modulation weights per-channel and δ denotes ReLU. The recalibrated feature is then computed as

$$\begin{aligned} \tilde{x_i} &= F_{scale}(x_i, a_i) = a_i x_i \\ \tilde{X} &= [\tilde{x}_1, \tilde{x}_2, ...\tilde{x}_C]. \end{aligned} \tag{10.6}$$

We proceed to discuss the dataset (Fig. 10.4).

10.3 Dataset

The FaceForensics++ dataset [40] comprises 1000 talking subjects, represented in 1000 real videos. Further, based on these 1000 real videos, 4×1000 adversarial examples have been generated by following four manipulation schemes.

1. **Faceswap** represents a graphic approach transferring a full face region from a source video to a target video. Using facial landmarks, a 3D template model employs blend-shapes to fit the transferred face. FaceSwap.[9]
2. **Deepfakes** has become the synonym for all face manipulations of all kind, it origins to FakeApp[10] and faceswap github.[11]
3. **Face2face** [48] is a facial reenactment system that transfers the expressions of a source video to a target video, while maintaining the identity of the target person. Based on an identity reconstruction, the whole video is being tracked to compute per frame the expression, rigid pose, and lighting parameters.
4. **Neuraltextures** [47] incorporates facial reenactment as an example for a *Neural-Textures*-based rendering approach. It uses the original video data to learn a neural texture of the target person, including a rendering network that has been trained with a photometric reconstruction loss in combination with an adversarial loss. Only the facial expression corresponding to the mouth region is being modified, i.e., the eye region stays unchanged.

10.4 Algorithms

We select three state-of-the-art 3D CNN methods, which have excelled in action recognition. We proceed to briefly describe them.

- **I3D** [10] incorporates sets of RGB frames as input. It replaces 2D convolutional layers of the original Inception model by 3D convolutions for spatio-temporal modeling and inflates pre-trained weights of the Inception model on ImageNet as its initial weight. Results showed that such inflation has the ability to improve 3D models.
- **3D ResNet** [23] and **3D ResNeXt** are inspired by I3D, both extending initial 2D ResNet and 2D ResNeXt to spatio-temporal dimension for action recognition. We note that deviating from the original ResNet-bottleneck block, the ResNeXt-block introduces group convolutions, which divide the feature maps into small groups. We also conducted experiments with the 3D ResNet modified with squeeze-excitation blocks and non-local block, and the 3D ResNeXt modified with non-local block to investigate the effect of using self attention on these networks.

Given the binary classification problem in this work, we replace the prediction layer in all networks with a single neuron layer, which outputs one scalar value. All three networks have been pre-trained on the large-scale human action dataset Kinetics-400. We inherit the weights in the neural network models and further fine-tune the networks on the FaceForensics++ dataset in all our experiments.

[9] https://github.com/MarekKowalski/FaceSwap/.

[10] https://www.fakeapp.com.

[11] https://github.com/deepfakes/faceswap.

We detect and crop the face region based on facial landmarks, which we detect in each frame using the method from Bulat and Tzimiropoulos [9]. Next, we enlarge the detected region by a factor of 1.3, in order to include pixels around the face region.

10.5 Experiments

We conduct experiments on the manipulation techniques listed above with the algorithms I3D, 3D ResNet and 3D ResNext aiming at training and detecting (a) all manipulation techniques, (b) each manipulation technique separately, as well as (c) cross-manipulation techniques. Toward this, we split train, test, and validation sets according to the protocol provided in the FaceForensics++ dataset.

We use PyTorch to implement our models. The three entire networks are trained end-to-end on 4 NVIDIA V100 GPUs. We set the learning rates to $1e^{-3}$. For training, I3D accepts videos of 64 frames with spatial dimension 224×224 as input. The size of input of 3D ResNet and 3D ResNeXt are 16 frames of spatial resolution 112×112. For testing, we split each video into short trunks, each of temporal size of 250 frames. The final score assigned to each test video is the average value of the scores of all trunks.

We also investigate the impact of two attention mechanisms on 3D ResNet, namely, Squeeze-Excitation blocks and Non-local blocks. In the case of the 3D ResNet with the Squeeze-Excitation (SE) blocks, the network is trained from scratch as the SE blocks are incorporated in the bottleneck modules themselves. Despite this addition not performing at par with the original 3D ResNet pre-trained on Kinetics, training is more stable and obtains superior results compared to a 3D ResNet that is trained on the dataset from scratch. Based on the limitations and advantages we observe for the 3D ResNet, we also investigate the impact of using the non-local block in the 3D ResNeXt, which outperform the other 3D architectures in most cases after this modification. We report in all experiments the true classification rates (TCR).

10.5.1 All Manipulation Techniques

Firstly we evaluate the detection accuracy of the three video CNNs (with and without attention), and compare the results to *image*-forgery detection algorithms. For the latter we have in particular the state-of-the-art XceptionNet [40], learning-based methods used in the forensic community for generic manipulation detection [8, 14], computer-generated vs. natural image detection [39] and face tampering detection [1]. Given the unbalanced classification problem in this experiment (number of fake videos being nearly four times the number of real videos), we use weighted cross-entropy loss, in order to reduce the effects of unbalanced data. We observe that among the unmodified 3D CNNs, the detection accuracy of I3D is the highest and it is also the most computationally intense. The performance of 3D ResNet improves with

Table 10.2 Detection of all four manipulation methods, LQ. TCR = True classification rate, DF = deepfakes, F2F = face2face, FS = face-swap, NT = neuraltextures

Algorithm	Train and test	TCR
Steg. Features + SVM [19]	FS, DF, F2F, NT	55.98
Cozzolino et al. [14]	FS, DF, F2F, NT	58.69
Bayar and Stamm [8]	FS, DF, F2F, NT	66.84
Rahmouni et al. [39]	FS, DF, F2F, NT	61.18
MesoNet [1]	FS, DF, F2F, NT	70.47
XceptionNet [13]	FS, DF, F2F, NT	81.0
I3D	FS, DF, F2F, NT	87.43
3D ResNet	FS, DF, F2F, NT	83.86
3D ResNet (w/o pre-training)	FS, DF, F2F, NT	54.96
3D ResNet (with SE)	FS, DF, F2F, NT	80.0
3D ResNet (with non-local)	FS, DF, F2F, NT	85.85
3D ResNeXt	FS, DF, F2F, NT	85.14
3D ResNeXt (with non-local)	FS, DF, F2F, NT	88.28

Table 10.3 AUC values of 3D ResNet and 3D ResNeXt endowed with attention

Algorithm	AUC
3D ResNet	0.82
3D ResNet (w/o pre-training)	0.51
3D ResNet (with SE)	0.72
3D ResNet (with non-local)	0.86
3D ResNeXt (with non-local)	0.91

the introduction of the non-local block. The lack of pre-training does hamper the performance of the 3D ResNet with the SE attention, however it performs significantly better than the vanilla 3D ResNet which was initialized with random weights. Interestingly, with the addition of the non-local block to the 3D-ResNeXt, its detection accuracy becomes the highest, surpassing I3D. Related results are depicted in Table 10.2. We present the receiver operating characteristic curves (ROC curves) in Fig. 10.5 and the area under the curve (AUC) in Table 10.3.

10.5.2 Single Manipulation Techniques

We proceed to investigate the performances of all algorithms, when trained and tested on single manipulation techniques. We report the TCRs in Table 10.4. Interestingly, here the video-based algorithms perform similarly as the best image-based algorithm. This can be due to the data-size pertaining to videos of a single manipulation

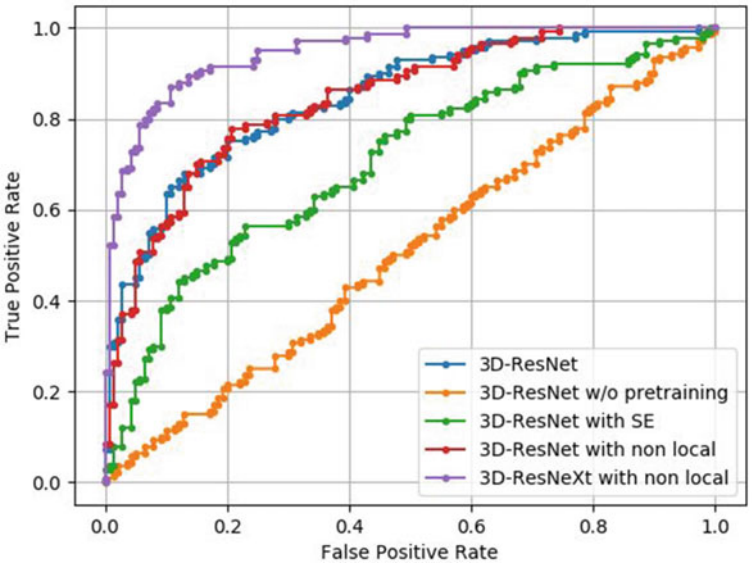


Fig. 10.5 ROC curves pertaining to 3D ResNet and 3D ResNext endowed with attention mechanisms for testing and training of all four manipulation methods

technique being smaller. I3D performed best among unmodified video-based methods. 3D ResNet with non-local block once again outperformed the pre-trained 3D ResNet and the 3D ResNet with SE attention outperformed the randomly initialized 3D ResNet that was trained from scratch. The performance of 3D ResNeXt also improved upon introduction of the non-local block, and in fact, it performed best among all video-based methods.

Our experiments suggest that all detection approaches are consistently utmost challenged on the GAN-based *neuraltextures*-approach. We note that *neuraltextures* trains a unique model for each video, which results in a higher variation of possible artifacts. While *deepfakes* similarly trains one model per video, a fixed post-processing pipeline is used, which is similar to the computer-based manipulation methods and thus has consistent artifacts that can be instrumental for deepfake detection.

10.5.3 Cross-Manipulation Techniques

In our third experiment, we train the 3D CNNs and the attention-endowed models with videos manipulated by 3 techniques, as well as the original (real) videos and proceed to test on the last remaining manipulation technique, as well as original videos. We show related results in Table 10.5. Naturally, this is the most challeng-

Table 10.4 Detection of each manipulation method individually, LQ. TCR = True classification rate, DF = deepfakes, F2F = face2face, FS = face-swap, NT = neuraltextures

Algorithm	DF	F2F	FS	NT
Steg. Features + SVM [19]	73.64	73.72	68.93	63.33
Cozzolino et al. [14]	85.45	67.88	73.79	78.00
Bayar and Stamm [8]	84.55	73.72	82.52	70.67
Rahmouni et al. [39]	85.45	64.23	56.31	60.07
MesoNet [1]	87.27	56.20	61.17	40.67
XceptionNet [13]	96.36	86.86	90.29	80.67
I3D	95.13	90.27	92.25	80.5
3D ResNet	91.81	89.6	88.75	73.5
3D ResNet (w/o pre-training)	58.80	73.60	59.20	56.50
3D ResNet (with SE)	81.70	77.00	75.90	66.25
3D ResNet (with non-local)	94.67	89.20	92.13	76.00
3D ResNeXt	93.36	86.06	92.50	80.50
3D ResNeXt (with non-local)	95.50	90.4	95.08	80.71

ing setting. At the same time, it is the most realistic one, because it is unlikely that knowledge on whether and how videos have been manipulated will be provided. Similar to the first experiment, we use weighted cross-entropy loss, in order to solve the unbalanced classification problem. For the detection algorithms, one of the more challenging settings in this experiment is when *faceswap* is the manipulation technique to be detected. We note that 3D ResNet with non-local block outperformed all other networks in this scenario.

While *face2face* and *faceswap* represent graphics-based approaches, *deepfakes* and *neuraltextures* are learning-based approaches. However, *faceswap* replaces the largest facial region in the target image and involves advanced blending and color correction algorithms to seamlessly superimpose source onto target. Hence the challenge might be due to the inherent dissimilarity of *faceswap* and learning-based

Table 10.5 Detection of cross-manipulation methods, LQ. TCR = True classification rate, DF = deepfakes, F2F = face2face, FS = face-swap, NT = neuraltextures, NL = non-local, scratch = w/o pre-training

Train	Test	3D	I3D	3D ResNeXt	3D ResNet (scratch)	3D ResNet (with SE)	3D ResNet (with NL)	3D ResNeXt (with NL)
FS, DF, F2F	NT	64.29	68.57	66.79	54.28	55.35	62.9	63.2
FS, DF, NT	F2F	74.29	70.71	68.93	51.0	53.5	68.2	69.1
FS, F2F, NT	DF	75.36	75.00	72.50	50.7	52.5	76.78	77.8
F2F, NT, DF	FS	59.64	57.14	55.71	50.3	53.5	68.2	65.71

Table 10.6 Number of parameters in 3D ResNet without and with attention

Algorithm	No. of parameters
3D ResNet	85,249,216
3D ResNet with SE	94,303,808
3D ResNet with non-local	93,647,040

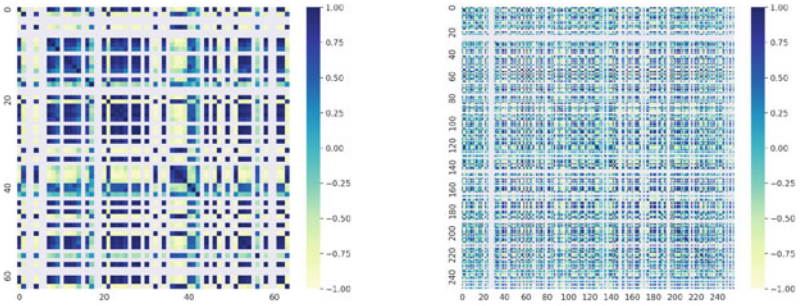
approaches, as well as due to the seamless blending between source and target, different than *face2face*.

We note that *humans* easily detected manipulations affectedResNet by *faceswap* and *deepfakes* and were more challenged by *face2face* and ultimately *neuraltextures* [40]. This is also reflected in the performance of 3D ResNet and 3D ResneXt with non-local block, which were most challenged by the videos manipulated by *neuraltextures*.

10.5.4 Effect of Attention in 3D ResNets

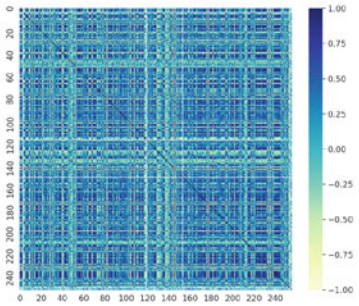
We here analyze the correlation matrices between two layers (at the same depth) for all the three variants of the 3D ResNet—the original 3D ResNet, the 3D ResNet with squeeze-excitation and the 3D ResNet with non-local block (Fig. 10.6). The high correlation observed in distinct patches in Fig. 10.6a indicates that the original 3D ResNet without attention possibly overfits to the data. The addition of squeeze-excitation (Fig. 10.6b) improves upon this and a further improvement is seen with the introduction of the non-local block in the 3D ResNet (Fig. 10.6c).

Both attention mechanisms, squeeze-excitation, and non-local block increase the number of parameters in the 3D ResNet by around 10% (Fig. 10.6), however when trained and tested on the whole dataset, we observe an improvement of 2% in the true classification rate in case of the model with non-local block (Table 10.2). We note that the 3D Resnet with SE attention could not be initialized with pre-trained Kinetics weights, so for a fair comparison, a 3D ResNet trained on the dataset from scratch was considered. Interestingly, without pre-trained weights, the vanilla 3D-ResNet is unable to converge its training in most cases and was underfitting. The training for the 3D ResNet with SE was more stable and yielded superior results over most experiments. It is also interesting to observe that *face2face* challenges 3D ResNet with non-local block more than the vanilla 3D ResNet. The exact reason behind this was not certain, however, as pointed out before, it was one of the more challenging scenarios for humans to detect as well [40]. In summary, 3D ResNet with the non-local block outperforms predominantly all other 3D ResNet variants (Table 10.2–10.5).



(a) 3D ResNet without attention

(b) 3D ResNet with SE



(c) 3D ResNet with non-local

Fig. 10.6 Correlation matrices for the 3D ResNets

10.5.5 Visualization of Pertinent Features in Deepfake Detection

We proceed to visualize features each of the 3D ResNet models are focusing on for detecting of deepfakes by Grad-CAM [43]. We note that Grad-CAM finds the final convolutional layer in a network and examines the gradient information flowing into that layer. The output of Grad-CAM is represented by a heat map visualization for a given class label, in our case deepfake detection. In particular, we visualize five frames from a *deepfake*-video in Fig. 10.7, for each of the three variants of 3D ResNet. Interestingly, we observe that 3D ResNet with both attention mechanisms focuses stronger on the central part of a face, as compared to the original 3D ResNet. It is also worth noting that the heat map for 3D ResNet with non-local block is located slightly higher than 3D Resnet with squeeze-excitation block, yielding the highest accuracy.

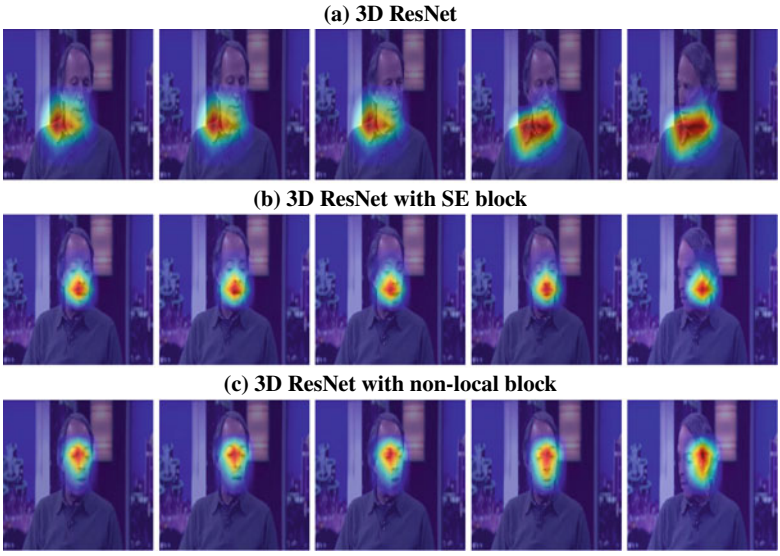


Fig. 10.7 Grad-CAM visualizations for the 3D ResNet models for the same video. The frames are taken from the same fake video with a time step of 24. Red represents higher probability of the region being manipulated

10.6 Conclusions

In this work we compared three state-of-the-art video-based CNN methods in detecting four deepfake-manipulation-techniques. The three tested methods included 3D ResNet, 3D ResNeXt and I3D, which we adapted from action recognition. In addition, we tested two attention mechanisms. Despite the pre-training of mentioned methods on the action recognition dataset Kinetics-400, the methods generalized very well to deepfake detection. Experimental results showed that 3D/video CNNs outperformed or performed at least similarly to image-based detection algorithms.

In addition, we observed that the incorporation of attention mechanisms in 3D CNNs improved related detection accuracy and were beneficial in placing focus of the models on areas of maximum manipulation in the forged videos.

Further, we noted a significant decrease in detection rates in the scenario, when we detected a manipulation technique not represented in the training set. One reason relates to the fact that networks lack an adaptation-ability to transfer learned knowledge from one domain (trained manipulation methods) to another domain (tested manipulation method). It is known that current machine learning models exhibit unpredictable and overly confident behavior outside of the training distribution.

Future work will involve the consideration of additional deepfake-techniques. Further, we plan to develop novel deepfake detection approaches, which place emphasis on appearance, motion as well as pixel-level-based generated noise, targeted to outsmart the improving generation and manipulation algorithms.

References

1. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS). IEEE (2018)
2. Agarwal S, El-Gaaly T, Farid H, Lim SN (2020) Detecting deep-fake videos from appearance and behavior. arXiv preprint arXiv:2004.14491
3. Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H (2019) Protecting world leaders against deep fakes. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 38–45 (2019)
4. Amerini I, Galteri L, Caldelli R, Del Bimbo A (2019) Deepfake video detection through optical flow based CNN. In: IEEE international conference on computer vision workshops (ICCVW), pp 1205–1207. https://doi.org/10.1109/ICCVW.2019.00152
5. Averbuch-Elor H, Cohen-Or D, Kopf J, Cohen MF (2017) Bringing portraits to life. ACM Trans Graph (TOG) 36(6):196
6. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
7. Baradel F, Wolf C, Mille J, Taylor GW (2018) Glimpse clouds: Human activity recognition from unstructured feature points. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 469–478
8. Bayar B, Stamm MC (2016) A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM workshop on information hiding and multimedia security. ACM, pp 5–10
9. Bulat A, Tzimiropoulos G (2017) How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: International conference on computer vision (ICCV)
10. Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset. In, IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6299–6308
11. Chaudhari S, Polatkan G, Ramanath R, Mithal V (2019) An attentive survey of attention models. arXiv preprint arXiv:1904.02874
12. Chesney R, Citron DK (2018) Deep fakes: a looming challenge for privacy, democracy, and national security. 107 California law review (2019, forthcoming); University of Texas Law. Public Law Research Paper 692:2018–2021
13. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
14. Cozzolino D, Poggi G, Verdoliva L (2017) Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM workshop on information hiding and multimedia security. ACM, pp 159–164
15. Cozzolino D, Rössler A, Thies J, Nießner M, Verdoliva L (2020) ID-Reveal: identity-aware deepfake video detection. arXiv preprint arXiv:2012.02512
16. Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) On the detection of digital face manipulation. In: IEEE/CVF conference on computer vision and pattern recognition, pp 5781–5790
17. Eichensehr K (2018) Don't believe it if you see it: deep fakes and distrust. Technology law; jotwell: the journal of things we like (lots), pp 1–2
18. Farid H (2011) Photo tampering throughout history. http://www.cs.dartmouth.edu/farid/research/digitaltampering
19. Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. IEEE Trans Inform Forensics Secur 7(3):868–882
20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems (NIPS), pp 2672–2680

21. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y (2018) Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927
22. Guarnera L, Giudice O, Battiato S (2020) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 666–667
23. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? In: IEEE conference on computer vision and pattern recognition (CVPR), pp 6546–6555
24. Hernandez-Ortega J, Tolosana R, Fierrez J, Morales A (2020) Deepfakeson-phys: deepfakes detection based on heart rate estimation. arXiv preprint arXiv:2010.00400
25. Hu J. Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7132–7141
26. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. arXiv preprint arXiv:1506.02025
27. Jiang L, Wu W, Li R, Qian C, Loy CC (2020) Deeperforensics-1.0: a large-scale dataset for real world face forgery detection. arXiv preprint arXiv:2001.03024
28. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4401–4410
29. Khalid H, Woo SS (2020) Oc-fakedect: classifying deepfakes using one-class variational autoencoder. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 656–657
30. Kim H, Carrido P, Tewari A, Xu W, Thies J, Niessner M, Pérez P, Richardt C, Zollhöfer M, Theobalt C (2018) Deep video portraits. ACM Trans Graph (TOG) 37(4):163
31. Korshunov P, Marcel S (2018) Speaker inconsistency detection in tampered video. In: 2018 26th European signal processing conference (EUSIPCO), pp 2375–2379. IEEE
32. Korshunov P, Marcel S (2019) Vulnerability assessment and detection of deepfake videos. In: The 12th IAPR international conference on biometrics (ICB), pp 1–6 (2019)
33. Li Y, Chang MC, Lyu S (2018) In ictu oculi: exposing AI generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877
34. Liu Z, Song G, Cai J, Cham TJ, Zhang J (2019) Conditional adversarial synthesis of 3D facial action units. Neurocomputing 355:200–208
35. Majumdar P, Agarwal A, Singh R, Vatsa M (2019) Evading face recognition via partial tampering of faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 11–20. https://doi.org/10.1109/CVPRW.2019.00008
36. Mnih V, Heess N, Graves A, Kavukcuoglu K (2014) Recurrent models of visual attention. arXiv preprint arXiv:1406.6247
37. Nadaraya EA (1964) On estimating regression. Theory Prob Appl 9(1):141–142
38. Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, Tran D (2018) Image transformer. In: International conference on machine learning (ICML), pp 4055–4064
39. Rahmouni N, Nozick V, Yamagishi J, Echizen I (2017) Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE workshop on information forensics and security (WIFS). IEEE
40. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to detect manipulated facial images. arXiv preprint arXiv:1901.08971
41. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3:1
42. Saito M, Matsumoto E, Saito S (2017) Temporal generative adversarial nets with singular value clipping. In: IEEE international conference on computer vision (ICCV), pp 2830–2839
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: IEEE international conference on computer vision (ICCV), pp 618–626. https://doi.org/10.1109/ICCV.2017.74
44. Silbey J, Hartzog W (2018) The upside of deep fakes. Md L Rev 78:960

45. Sun K, Liu H, Ye Q, Liu J, Gao Y, Shao L, Ji R (2021) Domain general face forgery detection by learning to weight. Proceedings of the AAAI conference on artificial intelligence 35:2638–2646
46. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing Obama: learning lip sync from audio. ACM Trans Graph (TOG) 36(4):95
47. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: image synthesis using neural textures. arXiv preprint arXiv:1904.12356
48. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M (2016) Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2387–2395
49. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inform Fusion 64:131–148
50. Trinh L, Tsang M, Rambhatla S, Liu Y (2021) Interpretable and trustworthy deepfake detection via dynamic prototypes. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), pp 1973–1983
51. Tulyakov S, Liu MY, Yang X, Kautz J (2018) MoCoGAN: decomposing motion and content for video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
52. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv preprint arXiv:1706.03762
53. Vinogradova K, Dibrov A, Myers G (2020) Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). Proceedings of the AAAI conference on artificial intelligence 34:13943–13944
54. Vondrick C, Pirsiavash H, Torralba A (2016) Generating videos with scene dynamics. In: Advances in neural information processing systems (NIPS) (2016)
55. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3156–3164
56. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
57. Wang Y, Bilinski P, Bremond F, Dantcheva A (2020) G3AN disentangling appearance and motion for video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 1–10. https://openaccess.thecvf.com/CVPR2020_search
58. Wang Y, Bilinski P, Bremond F, Dantcheva A (2020) ImaGINator conditional spatio-temporal gan for video generation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)
59. Wang Y, Bremond F, Dantcheva A (2021) InMoDeGAN interpretable motion decomposition generative adversarial network for video generation. arXiv preprint arXiv:2101.03049
60. Wang Y, Dantcheva A (2020) A video is worth more than 1000 lies. Comparing 3dcnn approaches for detecting deepfakes. In: FG'20, 15th IEEE international conference on automatic face and gesture recognition, May 18–22, 2020, Buenos Aires, Argentina
61. Wang Y, Dantcheva A, Bremond F (2018) From attributes to faces: a conditional generative adversarial network for face generation. In: International conference of the biometrics special interest group (BIOSIG), vol 17
62. Watson GS (1964) Smooth regression analysis. Indian J Stat, Ser A, Sankhyā, pp 359–372
63. Xu H, Ma Y, Liu H, Deb D, Liu H, Tang J, Jain A (2019) Adversarial attacks and defenses in images, graphs and text: a review. arXiv preprint arXiv:1909.08072