Demystifying Attention Mechanisms for Deepfake Detection

Abhijit Das, Srijan Das, Antitza Dantcheva

Demystifying Attention Mechanisms for Deepfake Detection

Abhijit Das[1], Srijan Das[2] and Antitza Dantcheva[3,4]
[1] Thapar Institute of Engineering & Technology, Patiala, India
[2] Stony Brook University, USA, [3] Inria, France, [4] Université Côte d'Azur, France

***Abstract*— Manipulated images and videos, i.e., deepfakes have become increasingly realistic due to the tremendous progress of deep learning methods. However, such manipulation has triggered social concerns, necessitating the introduction of robust and reliable methods for deepfake detection. In this work, we explore a set of attention mechanisms and adapt them for the task of deepfake detection. Generally, attention mechanisms in videos modulate the representation learned by a convolutional neural network (CNN) by focusing on the salient regions across space-time. In our scenario, we aim at learning discriminative features to take into account the temporal evolution of faces to spot manipulations. To this end, we address the two research questions 'How to use attention mechanisms?', and 'What type of attention is effective for the task of deepfake detection?' Towards answering these questions, we provide a detailed study and experiments 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.**

I. Introduction

Manipulated images and videos have become increasingly realistic and hence can pose serious security concerns and threats. Given that our society relies heavily on the ability to produce and exchange legitimate and trustworthy documents, *forged* images in driver's licences and passports can imply serious and far-reaching negative consequences on businesses, individuals, and political entities. While in the past multimedia manipulation was time consuming and costly, deep learning has well reduced costs, time and skill needed for realistic manipulation.

In the context of computer vision, creating *deepfakes* is an intriguing novel area of research [49], [47], [50]. However, deepfakes entail a number of challenges and threats, given that such manipulations can fabricate animations of subjects involved in actions that have not taken place, and such manipulated data can be spread 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 [37].

We can foresee deepfakes entailing the premise to inflict severe damage. Social threats [10], [16] can affect domains such as journalism and news media journalists[1][2][3][4]. In this context, we have two cases of concern. The first being *deepfakes being considered as real*, and the second relating to *real videos being considered as fake*.

Recent research on deepfake generation is able to forge short videos [42], [27], as well as to generate videos from a *single ID photo* [5]. In addition, fully synthesized *audio-video* images are able to replicate synchronous speech and lip movement [40] of a target person. Several deepfake-schemes have evolved till date. *Head puppetry* where the dynamics of a head from a source person are synthesized in a target person and *face swapping* where 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 are also performed in the first category. Currently such manipulations include subtle imperfections that can be detected by humans and, if trained well, by computer vision algorithms [30], [29], [3]. Towards detecting such attacks a number of multimedia forensics based detection strategies have been proposed [3], [36], [4], [14]. We note that such techniques have not been generalizable, i.e., were not able to provide a comprehensive solution against unseen manipulation techniques.

Therefore, generalizable deepfake detection remains a challenge. Existing detection methods [43], [53] have explored video classification in this context. Specifically, state-of-the-art networks successful in action recognition [51] such as I3D [9], 3D ResNet and 3D ResNext [20]. However, brute-force convolutional operations across space and time do not provide an optimal solution for deepfake detection, as 3D convolutions are too rigid to capture the subtle variations among original and generated fake videos. Therefore, an intuitive direction of research is towards applying attention mechanisms on top of aforementioned convolutional networks, as investigated in [35]. This raises a number of open questions related to attention mechanism for deepfake detection. Motivated by the above, in this paper we focus on attention mechanisms and related exploitation for deepfake detection. In particular, we investigate the following.

- What type of attention mechanism is adequate in deepfake detection?

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

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

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

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

- How should we exploit attention for deepfake detection? To be specific, where in the convolutional neural network the attention layers should be applied?
- As deepfake detection constitutes a video classification problem, how should we handle space and time, with respect to attention? Do we need to impose both spatial and temporal attention for deepfake detection? If so, then the next question is, should the network learn joint or dissociated spatio-temporal attention?

To answer the above questions, we revisit the popular deepfake detection methods, mostly inspired from the action recognition domain. Empirically, we find that spatio-temporal attention is ornamental for the task of deepfake detection. However, the next question is which type of attention is pertinent for this problem? To this end, our experimental analysis shows that self-attention mechanism when applied to all the convolutional layers is more effective in contrast to attention layers applied on top of the last convolutional layer. Interestingly, this observation is contradictory to the trends observed in action classification models. This suggests that deepfake detection is inherently different from classical action recognition and, hence requires specific approaches.
Finally, we also introspect this binary classification problem through vision transformers. This is mostly inspired from the chronology of experiments performed to answer the above questions. We see that vision transformers, unlike for action classification fail to outperform the convolutional counterpart for deepfake detection. But interestingly, the transformers even with small amount of training examples and no pre-training on a large dataset generalize over the training distribution rather than over-fitting as in convolutional networks in this domain.
Thus, in this paper, we discuss the best strategies to learn attention for detecting deepfakes. Through our extensive experiments, we show different settings for which a 3D convolutional network can be best exploited for deepfake detection. We also provide the end users a series of choices of networks (including vision transformers) for deepfake detection depending on their constraints. We believe that this experimental study will not only provide solution to best exploit the available resources for deepfake detection, but also shows potential research direction in this domain.

II. Related Work

Recent overview articles revisit the deepfake detection landscape [43], [53]. Generative adversarial networks (GANs) [18] have put on a direction for face manipulations including identity [24], [32], facial attributes [52], as well as facial expressions [31], [23], [48], [49], [50].

In the context of detecting such deepfakes a number of approaches were based on *image-analysis* [1], [34], *video analysis* [30], [3], [36] or jointly on audio and video analysis [28]. It is worth mentioning that some video analysis based manipulation approaches perform better than image-based ones, however such approaches are only applicable to particular kinds of attacks. For instance, many of them [30], [3] may fail, if the quality of the eye area is not sufficiently good or the synchronization between video and audio is not sufficiently natural [29]. Image-based approaches consists of general-purpose detectors, such as algorithm proposed by Fridrich and Kodovsky [17] and are hence applicable to both, steganalysis and facial reenactment video detection. Rahmouni et al. [33] presented an algorithm to detect computer-generated images, which was extended to detecting computer-manipulated images.

Agarwal *et al.* employed both, facial identity as well as behavioural biometrics information to produce the temporal component of videos to classify a video as real or fake [2]. Cozzolino *et al.* used temporal facial features to learn behaviour of a person [13]. 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 [19]. Khalid and Woo [26] posed deepfake detection as an anomaly detection problem and used variational autoencoder for detecting deepfakes. Montserra *et al.* proposed a network that exploits the advantages of both, convolutional, as well as recurrent network to classify the integrity of a video. Hernandez-Ortega [21] proposed a deepfake detection framework based on physiological measurement, namely heart rate using remote photoplethysmography (rPPG). Trinh *et al.* [45] utilized dynamic representations (i.e., prototypes) to explain deepfake temporal artifacts. Sun *et al.* [39] attempted to generalize forgery face detection by proposing a framework based on meta-learning.

Rössler *et al.* [34] 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.

III. Background: Existing Attention Mechanisms

In this section, we review attention mechanisms deemed relevant for deepfake detection. Attention mechanisms are generally designed to identify and focus on salient information to capture fine-grained information. Deepfake videos being accurate, possess subtle changes with respect to real videos. Therefore, attention mechanisms are instrumental in facilitating improved classification accuracy of a deepfake detector by enabling the model to focus on discriminative information. Firstly, we discuss attentional layers placed on top of convolutional networks. We begin by assuming that a video clip $V_c \in \mathcal{R}^{3\times H\times W}$ is sampled from the original video.

a) ***Squeeze and Excitation block****:* The Squeeze and Excitation (SE) block [22] boosts the representational power of a CNN by modelling inter-dependencies between channels of the features learnt by it. The SE block comprises of 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 a feature from a set of convolutional blocks $X \in R^{M\times N\times C}$ is represented as

X=[$x_1, x_2, ...x_C$], where $x_i \in R^{M \times N}$. The squeeze operation exploits global spatial information by squeezing X through global average pooling and creating a channel descriptor, $z \in R^C$ where i^{th} element of z is calculated as

$$z_i = F_{sq}(x_i) = \frac{1}{M \times N} \sum_{j=1}^{M} \sum_{k=1}^{N} x_i(j,k). \quad (1)$$

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.

$$a = F_{ex}(z, w_1, w_2) = \sigma(w_2 \delta(w_1 z)), \quad (2)$$

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

$$\begin{aligned} \tilde{x_i} &= F_{scale}(x_i, a_i) = a_i x_i \\ \tilde{X} &= [\tilde{x}_1, \tilde{x}_2, ...\tilde{x}_C]. \end{aligned} \quad (3)$$

*b) **Self-attention Block**:* In contrast to Squeeze-Excitation operation, self-attention block is characterized by computing the response at a position as a weighted sum of features at all positions in the input feature maps. The learning mechanism of attention weights for a space-time volume allow for modelling of long-term relationships.

Self-attention enables CNNs to capture long-range dependencies by activating relevant space-time pixels in the latent space through its own embeddings. Formally, in the context of CNNs, a self-attention operation in this scenario implemented via **non-local block** [46] is defined as

$$o_i = \frac{1}{C(x)} \sum_{\forall j} p(x_i, x_j) \;\; r(x_j), \quad (4)$$

where x and o denote the input and output features, 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

$$p(x_i, x_j) = e^{\alpha(x_i)^T \beta(x_j)}, \quad (5)$$

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.

*c) **Visual Transformer:**:* Self-attention mechanisms aimed at learning attention weights for a convolutional feature map are now being replaced with fully attentional networks. Therefore, we discuss and explore vision transformer [15], [44], [38] in our context. Deviating from former two attention mechanisms, we here aim at exploiting the texture of deepfake videos through vision transformers. We

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

Fig. 1. A sample frame from the FaceForensics++. From left to right: original source and target (small) images, deepfakes, face2face, faceswap, neuraltextures.

are particularly interested in video transformers, and thus focus on a state-of-the-art method, namely TimeSformer [7], which inputs patches (described as tokens) from a video clip. The patches of dimension $P \times P$ which are sub-parts of images are stacked across space and time. These patches while mapping into vectors are fed to a linear embedding to encode them into tokens. Then, these tokens are processed with self-attentional operations by projecting each of them into Keys ($K \in \mathcal{R}^{\mathcal{D}}$), Queries ($Q \in \mathcal{R}^{\mathcal{D}}$), and Values ($V \in \mathcal{R}^{\mathcal{D}}$). Finally, the modulated feature vector F is obtained by

$$F = softmax(\frac{Q \times K^T}{\sqrt{D}}) \times V. \quad (6)$$

This recombination of tokens in F indicates highlighting the pertinent tokens within a frame (for spatial attention), highlighting the pertinent temporal frame within a video clip (for temporal attention), and both (for spatio-temporal attention). The type of attention is implemented by invoking the above operation across the desired dimension. In our experiments, we explore both, joint spatio-temporal attention and dissociated spatial and temporal attention for detecting deepfakes. Note that dissociated spatial and temporal attention is implemented by applying temporal and then spatial attention, respectively.

IV. Dataset

The FaceForensics++ dataset [34] comprises of 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) **Face-swap** 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[5].
2) **Deepfakes** has become the synonym for all face manipulations of all kind, it origins to FakeApp[6] and faceswap github[7].
3) **Face2face** [42] 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

[5]https://github.com/MarekKowalski/FaceSwap/
[6]https://www.fakeapp.com
[7]https://github.com/deepfakes/faceswap

video is being tracked to compute per frame the expression, rigid pose, and lighting parameters.

4) **Neuraltextures** [41] incorporates facial reenactment as an example for a *NeuralTextures*-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.

V. Experiments

We select one state of the art 3D CNN method, namely 3D ResNet [20] as base network, which has excelled in action recognition. After each ResNet block we include above mentioned attention, in order to perform our experiments. The networks have been pre-trained on the large-scale human action dataset Kinetics-400 [25]. We inherit the weights in the neural network models and further fine-tune the networks on the FaceForensics++ dataset in all our experiments. We detect and crop the face region based on facial landmarks, which we detect in each frame using the method from Bulat and Tzimiropoulos [8]. Next, we enlarge the detected region by a factor of 1.3, in order to include pixels around the face region.

We conduct experiments on the manipulation techniques listed above: (a) all manipulation techniques, (b) each manipulation technique separately, as well as (c) cross-manipulation techniques. Towards this, we split train, test and validation sets according to the protocol provided in the FaceForensics++ dataset.

A. Implementation details

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 $1e^{-3}$. The size of input for 3D ResNet 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. For training TimeSformer, we use clips V_c of size 8×3×224×224, with frames sampled at a rate of 1/32. The patch size is 16 × 16 pixels. During inference, we sample a single temporal clip in the middle of the video. We use 3 spatial crops (top-left, center, bottom-right) from the temporal clip and obtain the final prediction by averaging the scores for these 3 crops. We note that the difference in the training/testing mechanism for 3D ResNet and TimeSformer but this is owing to their configuration for achieving the best performance. We report in all experiments the true classification rates (TCR).

B. All Manipulations Experiments

We evaluate the performance of 3D ResNet with or without attention, and also the recently popular visual transformers for the task of deepfake detection.

TABLE I

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 [17]	FS, DF, F2F, NT	55.98
Cozzolino *et al.* [12]	FS, DF, F2F, NT	58.69
Bayar and Stamm [6]	FS, DF, F2F, NT	66.84
Rahmouni *et al.* [33]	FS, DF, F2F, NT	61.18
MesoNet [1]	FS, DF, F2F, NT	70.47
XceptionNet [11]	FS, DF, F2F, NT	81.0
3D ResNet	FS, DF, F2F, NT	83.86
3D ResNet (with SE)	FS, DF, F2F, NT	80.0
3D ResNet (non-local 1,2,4)	FS, DF, F2F, NT	85.85
3D ResNet (non-local 4,4)	FS, DF, F2F, NT	81.79
3D ResNet (non-local 1,2,3,4)	FS, DF, F2F, NT	**86.72**
TimeSformer	FS, DF, F2F, NT	82.3

One of the challenges in deepfake detection is that the dataset is unbalanced i.e. the number of fake videos being nearly four times the number of real videos. To handle this issue, we use weighted cross-entropy loss to negate the biased scenario. The results are provided in Table I. We compare the results with image-forgery detection algorithms and the state-of-the-art such as XceptionNet [34], learning-based methods used in the forensic community for generic manipulation detection [12], [6], computer-generated vs. natural image detection [33] and face tampering detection [1].

After observing the experimental results, we conclude that the video-based algorithms perform similar to the image-based algorithm XceptionNet. This may be due to the smaller size of the training data. We employ different attention mechanism on 3D ResNet discussed in section III. We employ, SE based attention, non-local block on top of 3D convolutional network and finally also TimeSformer (a transformer type network). Further, we note that non-local attentional layers when placed on top of each block of 3D ResNet, outperform the pre-trained 3D ResNet, Transformer and the 3D ResNet with SE attention. We also note that Kinetics-based pretraining significantly boosts the detection performance.

We note that tampering of videos in the FaceForensics++ dataset is done either by replacing the largest facial region in the target image and advanced blending and color correction algorithms, or by learning-based manipulation models. Hence, seamlessly superimposing source onto target videos poses an inherent dissimilarity. Hence, capturing long-range dependencies as in the proposed non-local combination is highly beneficial in video processing for this scenario rather than previous settings proposed in [46]. However, it should be considered that the proposed setting is computationally expensive, as attention is employed at the lower layers, where the feature size is larger.

C. Single Manipulation Experiments

In this section, we investigate the performances of all baselines when trained and tested on single manipulation

TABLE II

ABLATION STUDY, SHOWCASING THE EFFECTIVENESS OF THE NON-LOCAL BLOCKS BASED ON ITS POSITION. (NON-LOCAL 4,4 IMPLIES 4 NON-LOCAL BLOCKS AFTER 4TH RESNET BLOCK, NON-LOCAL 1,2 IMPLIES 1 NON-LOCAL BLOCK AFTER 1ST AND 2ND RESNET BLOCK)

Network	Attention	DF	F2F	FS	NT
3D ResNet	Non-local 1,2	90.45	86.64	90.52	75.09
3D ResNet	Non-local 3,4	90.48	87.11	89.19	75.91
3D ResNet	Non-local 1,2,3	91.71	88.09	90,01	75.13
3D ResNet	Non-local 4,4	91.11	88.27	90.04	77.05
3D ResNet	Non-local 1,2,4	94.67	89.20	92.13	76.00
3D ResNet	Non-local 1,2,3,4	**95.16**	**91.25**	**94.11**	**78.29**

TABLE III

ABLATION STUDY, DETERMINING THE EFFECTIVENESS OF DIFFERENT ATTENTION TYPES (JOINT OR DISSOCIATED SPATIAL AND TEMPORAL) FOR DEEPFAKE DETECTION.

Network	Attention	DF	F2F	FS	NT
TimeSformer	Joint	**90.7**	**73.9**	**82.1**	67.5
TimeSformer	Dissociated	87.9	72.9	80.1	**69.3**

techniques. We report the TCRs in Table IV. The pattern of the results found to be very similar to the above discussed, where training and testing was performed on all manipulation techniques.

Our results suggest that the most challenging manipulation approach is the GAN-based *neuraltextures*-approach. It is to be noted that *neuraltextures* trains a unique model for each video, which results in a higher variation of possible artifacts, which can be reason for the low detection rates. We note that attention is beneficial also in this setting. In particular, while applying non-local based attention after each 3D ResNet block accounts for nearly 5% better detection rates. While *deepfakes* similarly train 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.

We perform ablation studies, in order (i) to determine the optimal position of the non-local block, and (ii) for the best employment of TimeSformer attention for the problem in hand, which is summarised in Table II and Table III, respectively. It can be concluded from Table II that in deepfake scenarios non-local block when employed as in the original paper[46] (i.e. 4 non-local blocks after the last ResNet block) under-performs compared to other presented configurations. We note that employing a single non-local block after each ResNet block outperforms all other representative baselines. Results of our ablation study, showing the importance of utilizing pre-trained weight is presented is Table V.

We proceed to tackle the question of handling space and time with respect to attention for deep fake. In SE a dissociated approach is followed, as squeeze 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

TABLE IV

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 [17]	73.64	73.72	68.93	63.33
Cozzolino *et al.* [12]	85.45	67.88	73.79	78.00
Bayar and Stamm [6]	84.55	73.72	82.52	70.67
Rahmouni *et al.* [33]	85.45	64.23	56.31	60.07
MesoNet [1]	87.27	56.20	61.17	40.67
XceptionNet [11]	**96.36**	86.86	90.29	**80.67**
3D ResNet	91.81	89.6	88.75	73.5
3D ResNet (SE)	81.70	77.00	75.90	66.25
3D ResNet (non-local 1,2,3,4)	95.16	**91.25**	**94.11**	78.29
TimeSformer	90.7	73.9	82.1	67.5

TABLE V

ABLATION STUDY, INDICATING THE PERTINENCE OF KINETICS PRETRAINING ON DEEPFAKE DETECTION. ATT INDICATES ATTENTION.

Network	Pre-train	DF	F2F	FS	NT
3D ResNet	×	58.80	73.60	59.20	56.50
3D ResNet	✓	91.81	89.6	88.75	73.5
3D ResNet (with att)	✓	**95.16**	**91.25**	**94.11**	**78.29**
TimeSformer	×	88.6	74.6	77.9	66.8
TimeSformer	✓	90.7	73.9	82.1	67.5

re-calibration weights. In Non-local with our setting, it is clear that the non-local blocks can capture long range spatial-temporal features, when trained jointly. For the experiments on TimeSformer based on joint and dissociated learning, we find that joint learning outperforms all scenarios except for neuraltextures, which is the most challenging. Therefore, it is clear that joint spatial-temporal attention mechanism is most effective in detecting deepfakes.

D. Cross-manipulation experiments

In our third set of experiments, we train our baselines and attention-empowered models with videos manipulated by 3 techniques, as well as with original videos and proceed to test on the remaining manipulation technique and original videos. We present related results in Table VI. We note that cross-manipulation constitutes the most challenging experiment setting. At the same time it is the scenario, which we can expect to encounter in practice, as it is unlikely that we will have knowledge of a possible manipulated technique.

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 outperforms all other networks in this setting. Among all manipulation techniques, face2face and faceswap represent graphics-based approaches, whereas deepfakes and neuraltextures are learning-based approaches. We have that in faceswap the full facial region in the target image is replaced by the source face image and involves advanced blending and color correction to accurately superimpose source onto target. Hence in this context the challenge is due to the

inherent dissimilarity of faceswap and the other manipulation techniques.

We note that *humans* can easily detect manipulations affected by *faceswap* and *deepfakes* and were more challenged by *face2face* and ultimately by *neuraltextures* [34]. This is also reflected in the performance of 3D ResNet with non-local block, which is mostly challenged by videos manipulated by *neuraltextures*. However, we have that in most scenarios, employment of attention mechanism is effective.

VI. Conclusions

In this work, we study attention mechanisms in the context of deepfake detection. Our results suggest that the incorporation of attention mechanisms improves the detection accuracy of deepfakes by placing the focus on artefacts in forged videos. We note that attention capturing long-term dependencies, namely self-attention via non-local blocks is best in our setting, when utilized in our proposed configuration. We conduct experiments in a cross-manipulation scenario, which remains the most challenging detection scenario with respect to detection rates. Specifically, reduced detection rates are more prominent for learned manipulation techniques such as *neuraltextures*, when not included in the training set. This suggests that current deepfake detection approaches lack in adapting to a distribution different from the training distribution, hence posses low generalization capabilities. While attention mechanism reduce the gap to some extent, generalizable deepfake detection remains an open challenge. Future work will involve blending different attention mechanisms to obtain more robust representation for deepfake detection.

References

[1] Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: 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, S.N.: Detecting deep-fake videos from appearance and behavior. arXiv preprint arXiv:2004.14491 (2020)

[3] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: 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.: Deepfake video detection through optical flow based cnn. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 0–0 (2019)

[5] Averbuch-Elor, H., Cohen-Or, D., Kopf, J., Cohen, M.F.: Bringing portraits to life. ACM Transactions on Graphics (TOG) **36**(6), 196 (2017)

[6] Bayar, B., Stamm, M.C.: 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, pp. 5–10. ACM (2016)

[7] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? CoRR **abs/2102.05095** (2021). URL `https://arxiv.org/abs/2102.05095`

[8] Bulat, A., Tzimiropoulos, G.: 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) (2017)

[9] Carreira, J., Zisserman, A.: 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 (2017)

[10] Chesney, R., Citron, D.K.: Deep fakes: A looming challenge for privacy, democracy, and national security. 107 california law review (2019, forthcoming); U of Texas Law. Public Law Research Paper (692), 2018–21 (2018)

[11] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258 (2017)

[12] Cozzolino, D., Poggi, G., Verdoliva, L.: 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, pp. 159–164. ACM (2017)

[13] Cozzolino, D., Rössler, A., Thies, J., Nießner, M., Verdoliva, L.: ID-Reveal: Identity-aware deepfake video detection. arXiv preprint arXiv:2012.02512 (2020)

[14] Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5781–5790 (2020)

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

[16] Eichensehr, K.: Don't believe it if you see it: Deep fakes and distrust. Jotwell: J. Things We Like p. 1 (2018)

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

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems (NIPS), pp. 2672–2680 (2014)

[19] Guarnera, L., Giudice, O., Battiato, S.: Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 666–667 (2020)

[20] Hara, K., Kataoka, H., Satoh, Y.: 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 (2018)

[21] Hernandez-Ortega, J., Tolosana, R., Fierrez, J., Morales, A.: Deepfakeson-phys: Deepfakes detection based on heart rate estimation. arXiv preprint arXiv:2010.00400 (2020)

[22] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)

[23] Jiang, L., Wu, W., Li, R., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real world face forgery detection. arXiv preprint arXiv:2001.03024 (2020)

[24] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410 (2019)

[25] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. CoRR **abs/1705.06950** (2017). URL `http://arxiv.org/abs/1705.06950`

[26] Khalid, H., Woo, S.S.: Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 656–657 (2020)

[27] Kim, H., Carrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (TOG) **37**(4), 163 (2018)

[28] Korshunov, P., Marcel, S.: Speaker inconsistency detection in tampered video. In: 2018 26th European Signal Processing Conference (EUSIPCO), pp. 2375–2379. IEEE (2018)

[29] Korshunov, P., Marcel, S.: Vulnerability assessment and detection of deepfake videos. In: The 12th IAPR International Conference on Biometrics (ICB), pp. 1–6 (2019)

[30] Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877 (2018)

[31] Liu, Z., Song, G., Cai, J., Cham, T.J., Zhang, J.: Conditional adversarial synthesis of 3D facial action units. Neurocomputing **355**, 200–208 (2019)

TABLE VI

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 ResNet	3D ResNet (scratch)	3D ResNet (with SE)	3D ResNet (with NL134)	3D ResNet (with NL1234)	Time-Sformer
FS, DF, F2F	NT	64.29	54.28	55.35	62.9	**65.2**	60.0
FS, DF, NT	F2F	74.29	51.0	53.5	68.2	**73.1**	62.5
FS, F2F, NT	DF	75.36	50.7	52.5	76.78	**77.8**	65.0
F2F, NT, DF	FS	59.64	50.3	53.5	**68.2**	65.71	53.6

[32] Majumdar, P., Agarwal, A., Singh, R., Vatsa, M.: Evading face recognition via partial tampering of faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 0–0 (2019)

[33] Rahmouni, N., Nozick, V., Yamagishi, J., Echizen, I.: Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE Workshop on Information Forensics and Security (WIFS). IEEE (2017)

[34] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. arXiv preprint arXiv:1901.08971 (2019)

[35] Roy, R., Joshi, I., Das, A., Dantcheva, A.: Comparing 3D CNN architectures and attention mechanisms for deepfake detection. In: C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, C. Busch (eds.) Handbook of Digital Face Manipulation and Detection. Springer International Publishing (2022)

[36] Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P.: Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) **3**, 1 (2019)

[37] Silbey, J., Hartzog, W.: The upside of deep fakes. Md. L. Rev. **78**, 960 (2018)

[38] Steiner, A., Kolesnikov, A., , Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270 (2021)

[39] Sun, K., Liu, H., Ye, Q., Liu, J., Gao, Y., Shao, L., Ji, R.: Domain general face forgery detection by learning to weight. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2638–2646 (2021)

[40] Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) **36**(4), 95 (2017)

[41] Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. arXiv preprint arXiv:1904.12356 (2019)

[42] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: 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 (2016)

[43] Tolosana, R., Romero-Tapiador, S., Fierrez, J., Vera-Rodriguez, R.: Deepfakes evolution: Analysis of facial regions and fake detection performance. arXiv preprint arXiv:2004.07532 (2020)

[44] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: Mlp-mixer: An all-mlp architecture for vision. arXiv preprint arXiv:2105.01601 (2021)

[45] Trinh, L., Tsang, M., Rambhatla, S., Liu, Y.: 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 (2021)

[46] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803 (2018)

[47] Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: G3AN: This video does not exist. Disentangling motion and appearance for video generation. arXiv preprint arXiv:1912.05523 (2019)

[48] Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: G3AN disentangling appearance and motion for video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5264–5273 (2020)

[49] Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: ImaGINator conditional spatio-temporal gan for video generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020)

[50] Wang, Y., Bremond, F., Dantcheva, A.: InMoDeGAN interpretable motion decomposition generative adversarial network for video generation. arXiv preprint arXiv:2101.03049 (2021)

[51] Wang, Y., Dantcheva, A.: 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. (2020)

[52] Wang, Y., Dantcheva, A., Bremond, F.: From attributes to faces: a conditional generative adversarial network for face generation. In: International Conference of the Biometrics Special Interest Group (BIOSIG), vol. 17 (2018)

[53] Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., Jain, A.: Adversarial attacks and defenses in images, graphs and text: A review. arXiv preprint arXiv:1909.08072 (2019)