Deepfake Caricatures: Using Artifact Amplification to Expose Doctoring

Anonymous ICCV submission

Paper ID 9040

Abstract

Deepfakes are posing a serious threat to the integrity of our digital society. Detection techniques are essential to combat the evolving abilities of deepfake generation methods. Crucially, developing tools that allow humans to recognize doctored videos at first glance is key to avoid the spread of misinformation. Here, we introduce a new deepfake detection framework that intelligently distorts fake videos to amplify unnatural deepfake artifacts. Our artifact amplification module is semi-supervised by human data and creates interpretable attention maps highlighting artifacts. These maps improve the performance of our classifier, and allow us to generate Deepfake Caricatures: a transformation of the original deepfake video where artifacts and subtle unnatural movements are amplified to improve human recognition. In short, the contributions of this work are two-fold: we show improved performance on several deepfake detection datasets with a model explicitly attending to artifacts of fake videos, and we provide a pipeline to create "Deepfake Caricatures", a defense mechanism against deepfakes to help people better identify if a video has been doctored. Our approach, based on a mixture of human and artificial supervision, aims to further the development of countermeasures against fake visual content, and grants humans the ability to make their own judgment when presented with dubious visual media.

1. Introduction

Deepfakes are the result of using deep networks to swap one face for another with the intention to fool a viewer. Research on deepfakes has behaved in an adversarial manner, where new developments in flagging of doctored videos is followed by more efficient and realistic advances in generation [37].

As of late, the research related to combating deepfake has been focused on developing strong detectors. This tends to overlook a secondary problem: the impact of fake stimuli as they are first viewed. Research has shown that fake information tends to generate cognitive biases, *even when viewers are told that the information is fake beforehand*. Research illustrating the fluency heuristic (automatic trust over information that has been seen before) [28] shows that once a piece of false information has been seen, it automatically impacts future decisions. What if we could automatically distort such fake information in a way that blocks its initial absorption?

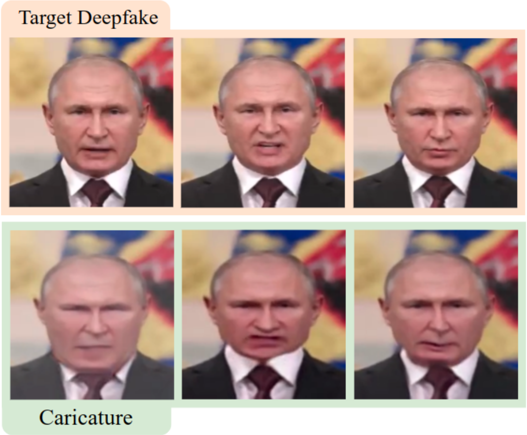

Figure 1. Examples frames from a fake original video of Vladimir Putin (top) and from the same video, after applying our Deepfake Caricatures process (bottom). Our method distorts the fake video to exacerbate artifacts and unnatural motion produced by the deepfake process. The jaw line, eye shape and general mouth movements are amplified in unnatural ways, evidencing the fake nature of the video. The caricatures are best experienced in video form. See supp. material for caricature videoclips.

Here, we develop an attention-based deepfake detection technique and explore for the first time the feasibility of exacerbating artifacts in deepfake videos to facilitate early detection from a human perspective (see Figure 1). As the quality of doctored videos becomes more impressive, too many generated fakes are indistinguishable from a genuine video to the human eye. We believe that it is crucial for humans to be able to detect, at first glance, if a video is doctored or not. This limits the spread of misinformation by stopping it at the source. Our approach works towards this goal by generating what we call **deepfake caricatures**: a

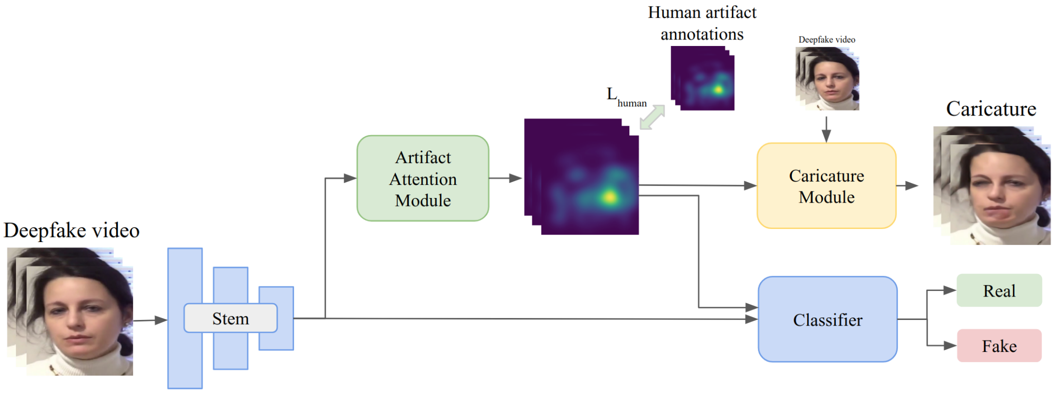

Figure 2. Our framework utilizes an artifact attention module that highlights defects, producing attention maps that improve classifier performance. Through supervision with human labels, we guide these artifact maps towards what humans consider fake, and generate caricatures: a transformation of the original videos where artifacts are amplified.

targeted distortion that reveals the fake nature of deepfakes, while rendering real videos virtually untouched. Our contributions are two-fold:

- We develop a novel deepfake detection technique that explicitly attends to artifacts in deepfakes. This is done through a novel **Artifact Attention Module** that learns to focus on defects, taking advantage of both human supervision and machine supervision (Grad-CAM [32] from successful pretrained models). We show our approach outperforms state of the art models on several benchmarks.
- We propose to generate **deepfake caricatures**, distorted versions of deepfakes that facilitate human recognition. We design an **Artifact Amplification Module** that uses our artifact attention to magnify unnatural movements in videos. We show in a controlled human study that our caricature module increases human performance in fake video detection by about 10% relative to baselines.

2. Related work

2.1. Deepfake face detection systems

Different categories of deepfake detection systems have been developed. A first category is based on variations in **network architectures**: Afchar et al. [4] propose an efficient low-complexity network aimed at detecting face swapping and face reenactment; Bayar and Stamm [5] develop a new convolutional layer that is better adapted to detecting forged images; and Nguyen et al. [26] use capsule networks in their detection pipeline. Several works [30, 13] utilize Recurrent Neural Networks (RNNs) with CNN features as inputs to classify videos as real or fake [24].

A second type of systems leverage **feature caveats** of fake faces and their generation process: Li et al. [18] proposes a detector that analyzes warping artifacts in deepfakes. Their solution does not require extensive deepfake training data, as large amounts of warping artifacts can be simulated by distorting any real face. Yang et al. [41] expose head pose inconsistencies in fake videos, using 3D head pose estimation to classify a video as real or fake. Ciftci et al. [9] show that biological signals are not preserved in fake videos, and they propose a multi-step pipeline that extracts these features and uses them for classification. Durall et al. [12] show that the Fourier transform of fake and real videos are significantly different using the amplitude spectrum of videos as input for a classifier. Matern et al. [21] use visual artifacts (errors in consistency, geometry, illumination, etc.) to improve fake detection, and show that this approach generates interpretable decisions.

A third category focuses on **attention mechanisms** that highlight specific parts of the input. Tolosana et al. [37] propose to segment frames into face regions (nose, eyes, etc.) and process each region with a separate model. Nguyen et al.[25] and Li et al. [16] train models to both classify the video and estimate a binary map highlighting the manipulated regions. Li et al. [17] focus on face swapping, and estimate the binary mask corresponding to the blending boundary in their detection pipeline. Finally, similar to our approach, Dang et al. [34] use an attention mechanism to guide the detector towards focusing on manipulated regions. Their method has no human supervision, however, and only focuses on traditional attention instead of actively amplifying the artifacts.

Our approach contributes to these three categories: we show that deepfakes are not robust to motion magnification and video distortion techniques; we propose to apply targeted artifact amplification by attending to specific parts of the video; and we develop a novel network that leverages this information as well as architectural and training improvements to surpass the state of the art in most benchmarks.

2.2. Motion amplification

We propose to amplify artifacts through an explicit manipulation module that is inspired from motion magnification techniques. Motion magnification consists of applying transformations to a given video to expose small movements and perturbations. Liu et al. [20] were the first to introduce this idea: their Lagrangian method estimates small motions and warps frames to expose them. Wadhwa et al. [40] developed a phase-based method that reduced noise and artifacts compared to previous approaches. Oh et al. [27] introduced a fully learned approach to amplify motion, based on a neural network that disentangles a frame's shape and texture before manipulation. To the best of our knowledge, we are the first to use a technique inspired from this work to amplify artifacts in fake videos.

2.3. Face caricatures

In caricatures, the most prominent or distinctive visual features of a face are exaggerated or distorted. Caricatures have been shown to be easier to recognize and remember [6, 22, 33, 39]. The proportions of the facial features seem to be more distinctive than the particular shape of a facial component [6]. Psychology research has shown that features encoded in memory correspond to deviations from a generic averaged face [6, 33], along with the proportion metric of facial features. As a caricature exaggerates distinct facial features, it draws attention to *facial regions that differ most from the norm*. Our deepfake caricatures emulate this phenomenon, and this increases a viewer's ability to recognize the video as fake.

3. Human attention maps of artifacts

The first step towards a model that incorporates human supervision into its pipeline is to collect human data. We designed several Amazon Mechanical Turk (AMT) tasks to obtain clean and useful human annotations. First, we utilized a simple selection interface to filter out easy-to-detect deepfake videos and enhance our video pool. We then showed the selected deepfakes to a new set of workers through a free-form annotation interface, where they were asked to manually paint over areas that appear distorted or manipulated.

Video selection We focused our data collection on videos which are difficult to spot as fake to begin with, as these are the most likely to yield non-trivial information about subtle artifacts. 1000 fake videos were selected from the DFDC dataset [11]: 500 videos which were difficult for humans to detect as fake, collected through a comparison-based human experiment (see supp. section Human experiments) and 500 videos that were hard to recognize as fake by the XceptionNet [29] detection model.

Task design: We designed a brush interface to allow participants to paint on the videos and highlight the areas that yield clues of fake videos (see supp. for data analysis). Workers were shown blocks of 100 3-second clips, and were asked to paint over the regions that appeared fake.

Results: we collected over 11K annotations for our 1000 videos. On average, each video had 22.6 sets of annotations, resulting in 4.1K sample points per video. For each video clip, we aggregate all annotation data and generate one 3D attention map. An anisotropic Gaussian kernel of size (20, 20, 6) in x, y and $time$ dimensions was first applied. We then normalize the attention map of each time frame to sum to one. Figure 3 illustrates some artifacts shown in fake videos, and Figure 4 shows some of the collected human maps, after averaging across participants and post-processing.

4. Deepfake detection with attention over artifacts

We present a framework that uses human-like attention to both detect and expose deepfakes to the human eye. Our goal is not only to detect fake videos with an algorithmic classifier, but also to make it easier for *humans* to detect fakes at a glance. By learning to attend to the most diagnostic regions, our model can detect fake videos as well as generate distorted videos (called *deepfake caricatures*) that are easier for humans to perceive as fake. Our deepfake detection framework contains three modules (see Figure 5):

- A **detector module** built on ResNet blocks that estimates whether a video is real or fake by considering the logit score of each frame;
- An **attention module** that extends a ResNet block by adding an attention layer, which extends the *self-attention* module and allows for supervision from human data;
- A **caricature module** that amplifies artifacts and unnatural motions with the guidance provided by the attention maps.

4.1. Detector module

Our basic detector module starts with a convolutional stem of three 3×3 convolution layers with stride of 2 in the first layer and 1 in the last two layers. In addition to ReLU activation functions, Mish [23] is also explored as a potential performance enhancement (MXResNet). A 3×3 max pooling layer with stride 2 follows the stem, and a sequence of ResNet blocks compose the rest of the network. We consider two types of ResNet blocks in our architecture: the common ResNet block of two 3×3 convolution layers and the bottleneck block of three convolution layers. We analyzed block sequence sizes of 18 and 34 in our model.

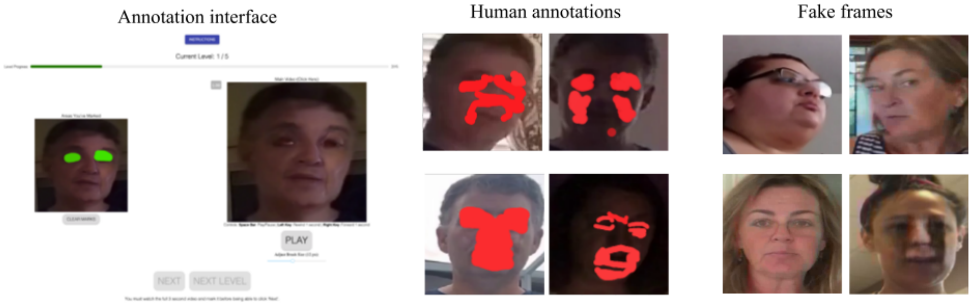

Figure 3. Our annotation interface (left) allows users to paint over particular artifacts or zones that appear fake. Our system tracks both the position and frame at which an annotation occurs. Users highlighted both large areas and more specific, semantically meaningful areas (e.g. eyebrows). Four examples of different types of artefacts shown in fake videos are illustrated on the right. Apart from the last one, all other types of artefacts seem to be subtle.

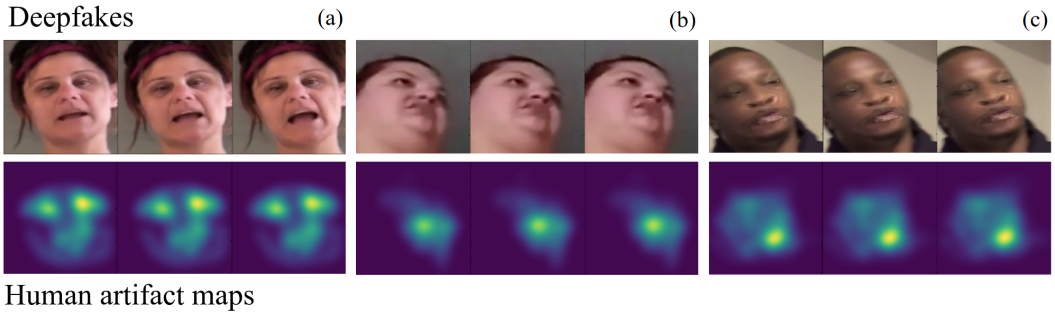

Figure 4. Examples of collected human maps. Humans consistently highlighted areas with abnormal artifacts. (a) the human maps correctly highlight the unnatural (but subtle) textures around eyes and mouth. (b) humans consistently marked the blurry mouth area as important to determine that the video is fake. (c) The artifacts over the cheek and mouth are highlighted; smooth textures around the left image side and the left eye (common in deepfakes) are marked, albeit with lower importance.

Cross entropy is used as the loss function. The output of the binary logit function is assigned to each frame as a detection score; we take the averaged score of the whole sequence as the prediction for the video clip.

4.2. Attention module

The attention module aims to incorporate human perceptual behavior into the detection pipeline and to guide the module towards informative regions which may not be locally connected. To do so, we first build an attention module which consists of a ResNet block and a self-attention layer. We then use the human data collected to supervise the generation of artifact attention maps.

Self-attention Built on Generative Adversarial Networks (SAGAN) [42], we extend the self-attention module by learning a downsampled attention map with a 2×2 max pooling layer. Given a feature map $\mathbf{x_i}$ generated from one image frame, the module learns to generate an attention map $\mathbf{a_i}$

$$\mathbf{a_i} = softmax((\mathbf{W_Q x_i})^{\mathbf{T}}(\mathbf{W_K x_i})). \tag{1}$$

The output of the attention layer y_i is

$$\mathbf{y_i} = \gamma \mathbf{a_i}^{\mathbf{T}}(\mathbf{W_V x_i}) + \mathbf{x_i}. \tag{2}$$

$\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V}$ are the learned weight matrices, and γ is a learnable scalar which balances between the learned attention feature and the original feature. Two 2×2 max-polling layers are applied to $\mathbf{W_K x_i}$ and $\mathbf{W_V x_i}$ to align the matrices.

Human attention We use the ground truth heat volumes generated from our crowd-sourced experiment as the supervision of the attention module. During training, the self-attention maps produced by our model are compared to the human attention maps, and specific losses are used to match the distributions of the two elements. We use Kullback-leibler divergence and Correlation Coefficient to force the learned attention maps to be more similar to the human attention data.

4.3. Caricature module

Our caricature module is built on the motion magnification network proposed in [27]. The attention maps guide this module: element-wise multiplication is utilized to fuse

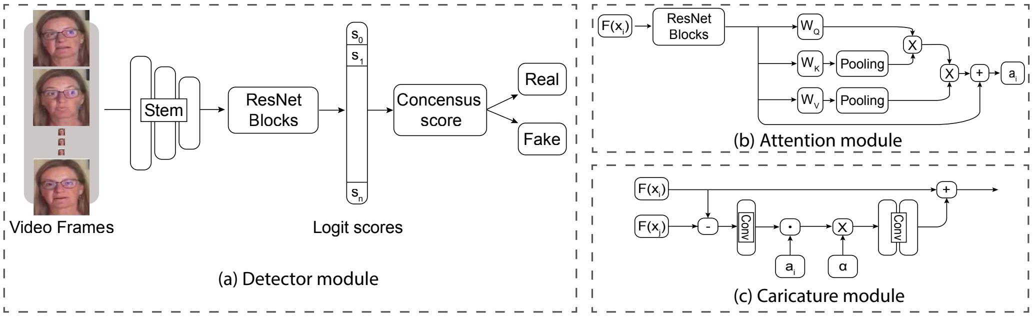

Figure 5. Three modules in our detection framework: (a) the detector module that estimates whether a video is real or fake, (b) the attention module that guides the network towards the key regions, and (c) the caricature module that amplifies unnatural artifacts. Consensus score is simply computed as the average mean in (a). $\times$ represents matrix multiplication and $\cdot$ represents element-wise multiplication. Refer to the text for details.

the information from the output of the motion magnification network with the information from the artifact attention module. The feature representation of the manipulated frame $\mathbf{F}'(\mathbf{x_i})$ is computed as

$$\mathbf{F}'(\mathbf{x_i}) = \mathbf{F}(\mathbf{x_i}) + \alpha(\mathbf{F}(\mathbf{x_j}) - \mathbf{F}(\mathbf{x_i})) \cdot \mathbf{a_i}, \qquad (3)$$

where $\mathbf{x_j}$ is an adjacent frame of $\mathbf{x_i}$ and α is the magnification factor used in the motion magnification network. By doing so, we target small inconsistent motions captured by the attention maps and amplify the corresponding artifacts in the generated caricature videos.

4.4. Detection and generation pipeline

The complete detection framework builds on the detector module with the ResNet blocks being replaced by the attention modules. Here we consider two variations: classification directly on the learned representation weighted by the attention maps and classification based on caricature videos (Figure 6. The later variation adds the caricature module to the pipeline, and feeds the caricature videos into another detector module to produce the final estimation. As a by-product, the **caricature videos generated by the caricature module with the guidance of attention maps may provide easy-to-detect characteristics for humans**.

5. Experiments

To evaluate our deepfake detection framework, we conduct extensive studies on available benchmark datasets, comparing to a large set of baseline models. We evaluate the model performance on the classification task, study whether a classifier trained on one dataset generalizes to others, and evaluate human detection accuracy on the generated caricatures.

5.1. Benchmarks

We evaluate models on four benchmarks:

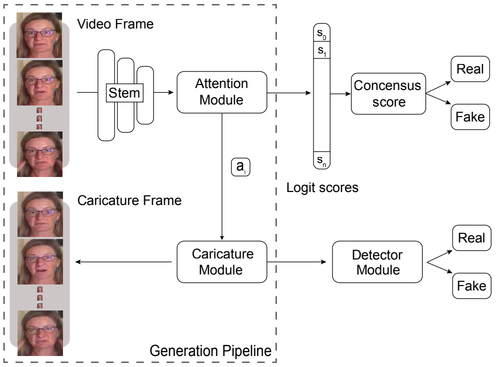

Figure 6. Diagram of the detection and generation pipeline, including detection at the caricature level. Adding the detector module after the caicature module allows us to test the performance of our detector after applying the caricature distortion.

The Deepfake Detection Challenge Dataset (DFDC) [11] contains 118,746 videos in the training set, including both the original and the manipulated videos. Two facial modification methods were used to generate the fake videos, however, no specific labels were given regarding the generation method. In this work, we report performance on the publicly available validation set of 400 labeled videos.

The FaceForensics++ dataset [29] contains more than 1.8M images generated from 5000 YouTube videos. It features four different facial manipulation methods, which are Deepfake [1], Face2Face [36], FaceSwap [2] and NeuralTextures [35], and covers a large variation in video resolution ranging from 1080p to 480p. We use 8151 images for training, 280 images for validation, and 280 for testing.

The Celeb-DF [19] dataset provides 5639 synthesized fake videos of celebrities generated from 590 real videos. Based on a tandem auto-encoder architecture, a synthesized method with special care of visual artifacts was used to generate high quality fake videos. In our experiment, we split the dataset into a training set of 6011 videos, and a validation set of 518 videos.

The YouTube-DF [3] dataset contains a large set of 1K deepfake videos from YouTube. It has been used to test the performance of state-of-the-art detection method on real-world videos. According to [3], the tested state-of-the-art methods perform poorly and did not generalize well on this set.

5.2. Baseline methods

We compare our model with the following baselines: the XceptionNet face detection model [29] and two variations of ResNet model, which operate on image frame and video (multiple frames).

XceptionNet: The XceptionNet was trained on the FaceForensics++ dataset and depending on the training set, we consider the Xception face detection model in our comparison, which achieved the best performance on the FaceForencis++ dataset [29].

Frame-based ResNet: We use ResNet [15] as a more general yet simple baseline for frame-based detection model and consider ResNet of block size 18 and 34 as two variations in the comparison.

Clip-based ResNet: Building on features learned from ImageNet, the inflation of pretrained 2D kernels into 3D has shown significant improvement on video-based tasks [8, 14, 38]. As a baseline of 3D CNN, We use a 3D ResNet model where the 3D kernels are inflated from ResNet18 pretrained on ImageNet [10]. We also include a ResNet3D pretrained on the Kinetics dataset [7] as another baseline.

5.3. Implementation details

Models were implemented in PyTorch, frame-based models pretrained on ImageNet when possible.

Data Preprocessing All datasets were downloaded at full resolution, except FaceForensics++, which was downloaded in the provided *c23* compressed format to avoid large storage costs. To streamline our training pipeline, we performed multi-person face extraction on all videos using a pretrained PyTorch implementation of FaceNet[31], with a minimum face size of 50 pixels and a face margin size of 100 pixels. For frames without a detected face, bounding box coordinates were produced via linear interpolation of the next available neighboring bounding boxes. To avoid abrupt motion artifacts (e.g. jittering) from face extraction, the 4 corner coordinates of the detected bounding boxes were temporally smoothed using hanning smoothing with a window size equal to 20% the video length. These face frames were resized to 360 × 360 pixels and saved as an MP4 file.

Learning and Optimization All models were optimized with RAdam (Rectified Adam) with LookAhead, an extension of the Adam optimizer that's informally referred to as the Ranger optimizer, with an initial learning rate of 0.001. Cosine annealing learning rate scheduling was applied with half period of 100 epochs. Batch size was chosen to maximally utilize available GPU memory and varied depending on the size of the model, but typically fell into the range of 32-64 videos per batch.

5.4. Impact of attention module on detection

We evaluate the performance of the proposed pipeline and baseline models. All benchmark datasets were used for training. We add details of the dataset splits in supp. section 2. Table 1, shows the binary classification accuracy of the baseline models on the testing datasets and the overall performance. Among all frame-based models, the detector module used in our framework achieved the best performance on each benchmark data apart from FaceForencis++. Interestingly, the clip-based models were mostly outperformed by the majority of frame-based models. While the large-scale Kinetics dataset improves learning in videos, further analysis is needed to determine the reason behind its under-performance on the other benchmarks. We hypothesize that frame-based models are more accurate at detecting fine grained defects in a single image, and these visual defects might be more effective than the motion information leveraged by 3D networks. By combining their predictions across multiple frames, the 2D models emulate an ensemble of image artifact detectors. Among the baselines, ResNet34 shows the highest performance with an overall classification accuracy of 95.68%.

We then tested whether the **attention mechanism could improve the detection**. We first enabled the self-attention layers in the ResNet blocks and experimented with two variations of ResNet18 and ResNet34, labelled as *SAResNet18* and *SAResNet34* in Table 1. The performance further improves when the human ground truth data was used as supervision (labelled as *Sup SAResNet34*), achieved the best overall accuracy of 97.06% among all tested models. Similarly, we evaluated the **performance of the caricature module** by running the simple detector module on distorted videos. The caricature module was pretrained on the motion magnification dataset [27] and we compared to the model where the pretrained network weights were frozen (*F-CariResNet*). Lastly, we tested the full pipeline using ResNet34 as the detector with supervised attention module (*Full-ResNet34*). The caricature module has the frozen pretrained weights. Using pretrained distortion weights without updates seems to improve the detection when compared to the performance of the un-frozen model.

5.5. Generalization to other datasets

To study the fact that deepfake classifiers do not seem to generalize well when presented with unseen data, we trained our module on one benchmark dataset and tested on the remaining three datasets. We varied the module connec-

	Model	DFDC	FF++	Celeb	YouTube	Overall
Frame-based	XceptionNet	58.0	**98.93**	67.57	50.0	68.63
	ResNet18	93.25	82.86	96.91	100	93.26
	ResNet34	96.00	90.00	96.72	100	95.68
Clip-based	Detector module	96.00	88.57	97.30	100	95.47
	ResNet3D (ImageNet)	90.73	78.57	90.73	100	90.01
	ResNet3D (Kinetics)	88.75	96.07	92.66	92.85	92.58
With attention (ours)	SAResNet18	92.25	94.29	92.09	96.43	93.77
	SAResNet34	95.5	94.29	**98.46**	100.0	**97.06**
	Sup SAResNet34	**96.25**	93.93	98.07	100.0	**97.06**
With caricature (ours)	CariResNet18	84.75	90.0	94.40	82.14	87.82
	F-CariResNet18	94.00	88.57	93.05	100	93.90
Full pipeline (ours)	Full-ResNet34	77.0	97.86	80.12	78.57	83.39

Table 1. Detection performance results. We report the classification accuracy on four tested benchmark datasets. Chance is 50%. Among all models, the highest accuracy is highlighted in **bold**. Three frame-based models, two clip-based models and the detector module in our framework provide baseline performance. *With attention* shows the results when the attention module is combined with the detector. *Sup* stands for human supervision. *With caricature* shows the results when all three modules are enabled. The caricature model was pretrained on the motion magnification dataset. F-CariResNet corresponds to a version of the network where the magnification weights are frozen.

	Model	FF++	Celeb	YouTube
[DFDC]	RN34 (Baseline)	57.50	82.24	**89.29**
	MXRN34 (Baseline)	65.71	78.76	78.57
	With Attention (Ours)	**68.21**	**82.63**	67.86
	Full pipeline (Ours)	61.07	82.05	71.43

	Model	DFDC	Celeb	YouTube
[FF++]	RN34 (Baseline)	52.25	**70.84**	**64.28**
	MXRN34 (Baseline)	50.0	65.63	50.0
	With attention (Ours)	**61.0**	70.08	**64.28**
	Full pipeline (Ours)	52.0	65.64	60.71

	Model	DFDC	FF++	YouTube
[Celeb]	RN34 (Baseline)	56.75	8.57	53.57
	MXRN34 (Baseline)	55.75	33.57	67.86
	With attention (Ours)	55.50	12.14	67.85
	Full pipeline (Ours)	**58.0**	**39.64**	**82.14**

Table 2. Model generalisation. Models were only trained on one dataset and tested on the other benchmarks. Results trained on DFDC are shown in (a), FF++ in (b), and Celeb in (c).

tions in the framework and both the attention and caricature modules improve the generalizability of baseline classifiers (Table 2).

5.6. Human detection results on deepfake caricatures

We run an AMT user study with 400 videos from the DFDC benchmark test set (200 fake and 200 real videos). People were asked to respond whether each video was real or fake. Videos were approximately 10 seconds long and the first 3 seconds were mandatory to watch. We collected 20 responses per video. Additional analyses and visualizations are in supp. materials section 2.2. We tested *three baseline conditions*: (1) **Original**: original videos, without added distortion (no caricature); (2) **Self-attention**: caricature videos guided by the learned self-attention maps only, without any human supervision. This condition is comparable to an ablation of the human loss on the attention maps; (3) **Uniform caricature**: caricature videos generated using the original motion magnification method [27] which distorts the video uniformly. We tested *two model caricatures* conditions: (1) **Human supervision**: caricature videos guided by the attention maps supervised by human data; (2) **Grad-CAM**: caricature videos guided by the Grad-CAM maps of our full model.

Table 3 present the results according to signal detection theory, focusing on the fake videos. Results show our model caricatures do not necessarily help to detect real videos in comparison to baselines caricatures. This is not surprising given the distorted features in real videos tend to be distributed randomly. Importantly, Grad-CAM generated caricatures on the deepfake videos significantly increase human detection of deepfakes.

As shown in Table 3, the Grad-CAM caricatures have the highest HIT score, with **71**% correct responses on the fake videos (which is significantly different than the other conditions, with $t(199) > 5$ and $p < .0001$). Grad-CAM caricatures elicited only **26**% of false alarms (FA, responding *fake* on a real video) and a d-prime (measure of sensitivity) of **1.2**. Importantly, the decision criterion C near zero indicates that in the Grad-CAM caricature condition, participants do not have a biased response towards fake or real, which is the best scenario for scaling up our method. As expected, this is in sharp contrast with the original video condition (no caricature): people cannot discriminate well even when they can scrutinize the videos, which video is real or fake, and have a tendency to respond *real* more often. The results also emphasize the importance of following our

Fake	Original (B)	Uniform (B)	Self-Attention (B)	Human-supervision	**Grad-CAM**
	No caricature	Caricature	Caricature	Caricature	Caricature
HIT	0.42	0.60	0.62	0.60	**0.71**
FA	0.22	0.40	0.20	0.18	**0.26**
d-prime	0.57	0.50	1.14	1.17	**1.20**
C	0.48	0.00	0.27	0.33	**0.04**

Table 3. The results are presented according to the signal detection theory: HIT and False Alarm (FA) rates of detecting a fake video. d-prime measures the sensitivity in detecting that a video is a **fake**; C is the decision criterion, or decision bias.

guided caricature generation process instead of a uniform artifact amplification, which creates caricatures that distorts human perception (very low d-prime of 0.5). While future work will focus on creating more efficient caricatures with a higher d-prime sensitivity, and little response bias, our results serve as a proof of concept that shows that deepfake caricatures improve human detection with minimal bias and impact on false alarms.

6. Conclusion

Many deepfakes are indistinguishable from a genuine video to the human eye. This can create individual and societal harms that can be prevented or mitigated with anti-fakes techniques. Here we capitalize on a novel framework which allows people to become aware of a fake by simply looking at a video caricature version. We show that our novel technique can generate a “caricature” of a video, a first of its kind approach, that augments the capability of humans to detect fakes. Our work establishes the feasibility of building relevant caricatures that impact human perception.

Importantly, one positive outcome of our work involves increasing trustworthiness and integrity of video-based media, by developing a robust automated response to combat fake news and viral threats on their initial presentations. Our caricature approach enhances the human ability to make judgments on videos, and stops the fake information before it can be absorbed. We believe that it is crucial for humans to be able to detect by themselves if a video is doctored or not, limiting the spread of misinformation by cutting it at its source. Caricatures are also known to be more memorable, which would help people remember that a particular video was fake. As is common in the field, our networks could be reverse engineered to produce deepfakes that are robust to our type of detection. It is important to maintain our method updated and trained on the latest datasets to mitigate the performance drop on novel generation techniques. It is important to note that our human-centric approach is less affected by novel generation techniques unless they are specifically tailored to be immune to the caricature generation system. For future general deepfake generation methods, our caricatures should still be able to give humans an edge at the moment of determining if a video is fake or not.

References

[1] Deepfake. https://github.com/deepfakes/faceswap, 2020. 5

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap/, 2020. 5

[3] Youtubedf. http://deepfake-detection.dessa.com/projects, 2020. 6

[4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 2

[5] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016. 2

[6] Philip J Benson and David I Perrett. Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology*, 3(1):105–135, 1991. 3

[7] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 6

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[9] Umur Aybars Ciftci and Ilke Demir. Fakecatcher: Detection of synthetic portrait videos using biological signals. *arXiv preprint arXiv:1901.02212*, 2019. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 3, 5

[12] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019. 2

[13] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018. 2

[14] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[16] Jia Li, Tong Shen, Wei Zhang, Hui Ren, Dan Zeng, and Tao Mei. Zooming into face forensics: A pixel-level analysis. *arXiv preprint arXiv:1912.05790*, 2019. 2

[17] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. *arXiv preprint arXiv:1912.13458*, 2019. 2

[18] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts, 2018. 2

[19] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 2019. 5

[20] Ce Liu, Antonio Torralba, William T Freeman, Frédo Durand, and Edward H Adelson. Motion magnification. *ACM transactions on graphics (TOG)*, 24(3):519–526, 2005. 3

[21] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019. 2

[22] Robert Mauro and Michael Kubovy. Caricature and face recognition. *Memory & Cognition*, 20(4):433–440, 1992. 3

[23] Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 2019. 3

[24] Daniel Mas Montserrat, Hanxiang Hao, SK Yarlagadda, Sriram Baireddy, Ruiting Shao, János Horváth, Emily Bartusiak, Justin Yang, David Güera, Fengqing Zhu, et al. Deepfakes detection with automatic face weighting. *arXiv preprint arXiv:2004.12027*, 2020. 2

[25] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. 2

[26] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019. 2

[27] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 633–648, 2018. 3, 4, 6, 7

[28] Gordon Pennycook, Tyrone D Cannon, and David G Rand. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865, 2018. 1

[29] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 3, 5, 6

[30] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3:1. 2

[31] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[33] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006. 3

[34] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation. *arXiv preprint arXiv:1910.01717*, 2019. 2

[35] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 5

[36] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 5

[37] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. Deepfakes evolution: Analysis of facial regions and fake detection performance, 2020. 1, 2

[38] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6

[39] Barbara Tversky and Daphna Baratz. Memory for faces: Are caricatures better than photographs? *Memory & cognition*, 13(1):45–49, 1985. 3

[40] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Phase-based video motion processing. *ACM Trans. Graph. (Proceedings SIGGRAPH 2013)*, 32(4), 2013. 3

[41] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses, 2018. 2

[42] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 4