Deepfake Caricatures: Amplifying Artifacts with Motion Magnification

Camilo Fosco*
MIT
camilolu@mit.edu

Alex Andonian*
MIT
andonian@mit.edu

Emilie Josephs
MIT
ejosephs@mit.edu

Allen Lee
MIT
allenlee@mit.edu

Xi Wang
MIT
nicole.xiwang@gmail.com

Aude Oliva
MIT
oliva@mit.edu

Abstract

Deepfakes are posing a serious threat to the integrity of our digital society by fueling the spread of misinformation. It is essential to develop detection techniques which can combat the evolving abilities of deepfake generation methods. One approach is to construct tools that allow humans to recognize doctored videos at first glance. Here, we introduce a new deepfake detection framework that intelligently distorts fake videos to amplify their artifacts. Specifically, we introduce a semi-supervised artifact attention module, which creates interpretable attention maps that highlight video artifacts based on human data. These maps make two contributions. First, they improve the performance of a deepfake detection classifier: we show higher accuracy across several deepfake detection datasets when the model explicitly attends to artifacts in fake videos. Second, the artifact attention maps allow us to generate novel "Deepfake Caricatures": a transformation of the original deepfake video where artifacts and subtle unnatural movements are exacerbated to improve human recognition. These caricatures are generated by a dedicated module that amplifies motion on detected artifacts by modifying the difference in the representation of consecutive frames, and leaves real videos untouched. 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 neural networks on images or videos to swap one face for another with the intention to fool a viewer. Research on deepfakes typically proceeds in an adversarial manner, where new developments in flagging of doctored videos are followed by more efficient and realistic advances in generation [40].

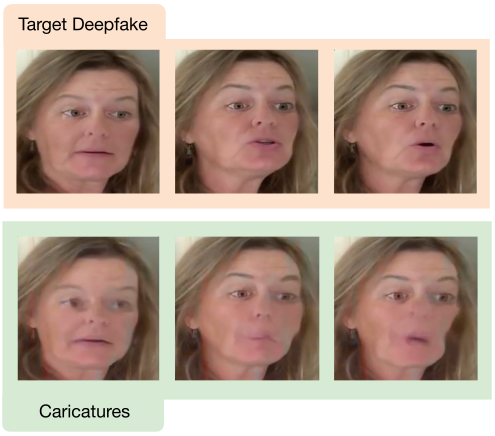


Figure 1. Example frames of a deepfake video (top row) and our deepfake caricature output (bottom row). Our method distorts the fake video to exacerbate artifacts and unnatural motion produced by the deepfake process. The jaw line, eye shape and general mouth movements are amplified in unnatural ways,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

As of late, the research related to combating deepfakes has been focused on developing strong detectors to flag false information. However, this approach overlooks the cognitive impact of exposure to fake stimuli. Research has shown that fake information tends to generate cognitive biases, *even when viewers are told that the information is fake beforehand.* For example, research illustrating the fluency heuristic (automatic trust over information that has been seen before) [31] shows that once a piece of false information has been seen, it automatically impacts future decisions. What if we could preemptively distort such fake information in a way that blocks its initial absorption?

Here, we develop an attention-based deepfake detection

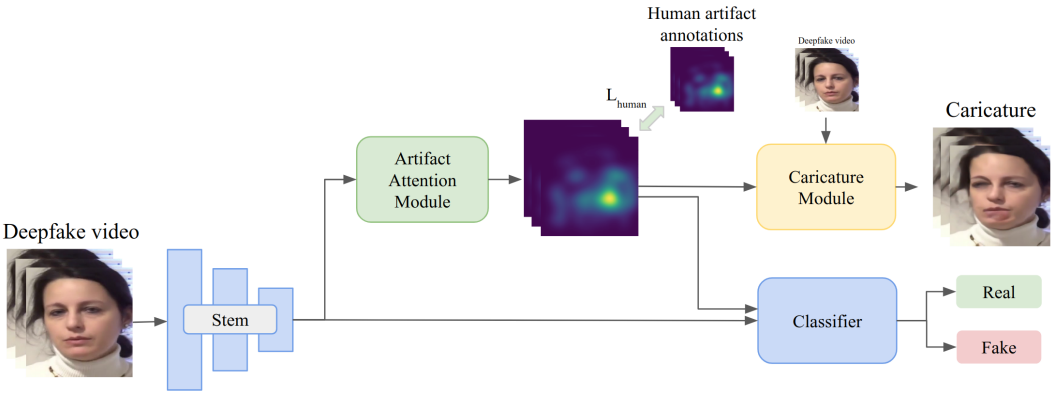


Figure 2. Our framework utilizes an artifact attention module that highlights defects, producing attention maps that improve classifier performance. Through supervision with human labels, we guide these artifact maps towards what humans consider fake, and generate caricatures: a transformation of the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

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 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 from successful pretrained models (e.g. Grad-CAM [35]). 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 detection systems

Different kinds of approaches have been explored for deepfake detection. One approach focuses 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. [29] use capsule networks in their detection pipeline. Several works [15, 33] utilize Recurrent Neural Networks (RNNs) with CNN features as inputs to classify videos as real or fake [27].

A second approach has been to leverage **feature caveats** inherent to fake faces and their generation process: Li et al. [20] propose 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. [44] expose head pose inconsistencies in fake videos, using 3D head pose estimation to classify a video as real or fake. Ciftci et al. [11] 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. [14] 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. [24] use visual artifacts (errors in consistency, geometry, illumination, etc.) to improve fake detection, and show that this approach generates interpretable decisions.

Finally, some approaches employ **attention mechanisms** that highlight specific parts of the input. Tolosana et al. [40] propose to segment frames into face regions (nose, eyes, etc.) and process each region with a separate model. Nguyen et al. [28] and Li et al. [18] train models to both classify the video and estimate a binary map highlighting the manipulated regions. Li et al. [19] 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. [37] use an attention mecha-

nism to guide the detector towards focusing on manipulated regions. Their method has no human supervision, however, and only focuses on traditional attention mechanisms instead of actively amplifying the artifacts.

Our pipeline makes advances to all three of these approaches: we combine targeted attention to specific parts of the video (attention mechanisms) with video distortion to highlight artifacts in deepfake videos (feature caveats). We then develop a novel network that leverages this information to surpass the state of the art in deepfake detection on most benchmarks, and develop a novel pipeline to amplify these artifacts, thereby improving human detection rates.

2.2. Motion amplification

We propose to amplify artifacts through a manipulation module targeted to specific locations, inspired by motion magnification techniques. Motion magnification consists of applying transformations to a given video to expose small movements and perturbations. Liu et al. [22] were the first to introduce this idea: their Lagrangian method estimates small motions and warps frames to expose them. Wadhwa et al. [43] developed a phase-based method that reduced noise and artifacts compared to previous approaches. Oh et al. [30] 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 the art style known as "caricature", prominent or distinctive features of a face are exaggerated or distorted in a way that makes them easier to recognize and remember [7,25,36,42]. Psychology research has shown that faces are encoded in memory based on their deviations from a generic averaged face [7,36], and that the proportions of facial features are at least as important as the particular shape of a facial component for distinguishing among faces [7]. Caricatures leverage these findings by exaggerating distinct facial features and the proportions among them, drawing attention to *facial regions that differ most from the norm*. Our deepfake caricatures emulate this phenomenon by making distortions in facial features and proportions more visible, increasing a viewer's ability to recognize the video as fake.

3. Human attention maps of artifacts

To create a model that incorporates human supervision into its pipeline, we first collected 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 create a video pool of deepfakes that are difficult for untrained humans to identify. We then showed the selected deepfakes to a new set of workers, who annotated areas of the videos that appeared distorted or manipulated.

Video selection Data collection was focused 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. We selected 500 videos from the DFDC dataset [13] which were difficult for humans to detect as fake, collected through a comparison-based human experiment (see supp. section Human experiments). We supplemented these with 500 additional videos from the DFDC dataset that were hard to recognize as fake by the XceptionNet [32] detection model, for a total of 1000 fake videos.

Video annotation We designed a brush interface to allow participants to paint on the videos and highlight the areas that appeared unnatural in the 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 combines self attention and human-based 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 main model, dubbed *CariNet*, contains three main modules (see Figure 2):

- A **Classifier module** built as a set of ResNet blocks with self-attention, where the self-attention is modulated by the output of the artifact attention module. The module outputs a Real or Fake classification for the input video.
- An **artifact attention module** that outputs heatmaps indicating the potential location of artifacts in each input frame. The model is supervised with ground truth artifact maps collected from our crowdsourced experiment. The heatmaps are fed to the classifier module,

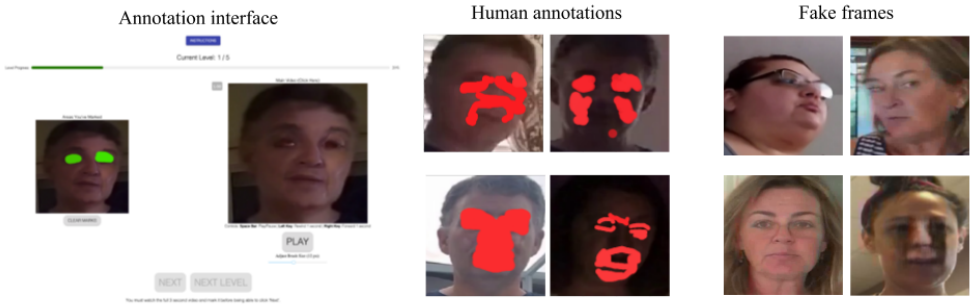


Figure 3. Our annotation interface (left) allows users to paint over particular artifacts or zones that appear fake. Our system tracks both the position and frame at which an annotation occurs. Users highlighted both large areas and more specific, semantically meaningful areas (e.g. eyebrows). Four examples 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.

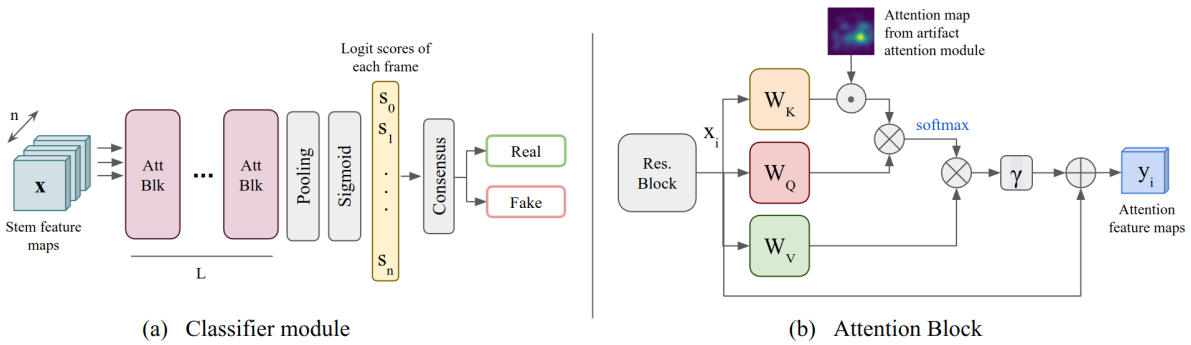
where they modulate the keys of the internal self attention layers, and to the Caricature module, where they control the spatial impact of the distortion.

- A **caricature module** that amplifies artifacts and unnatural motions with the guidance provided by the attention maps.

4.1. Classifier module

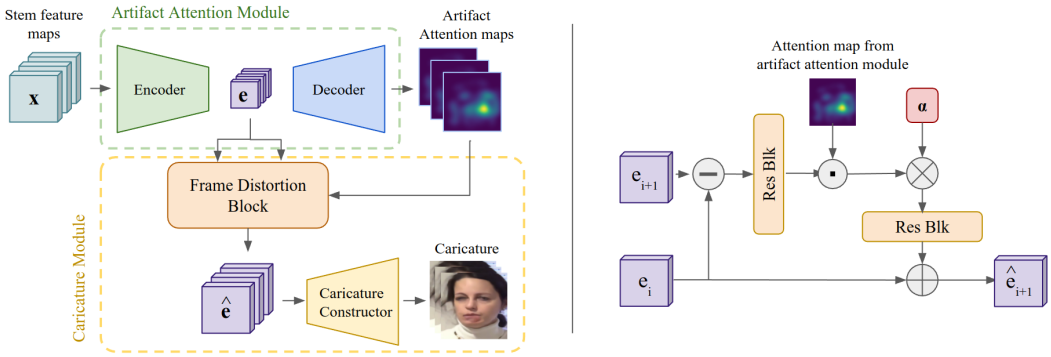
Our Classifier module (Figure 5) attempts to detect if the input is real or fake, and draws from the information provided by the artifact attention module to attend to key parts of the input. It is composed of a convolutional stem and a main trunk of Attention Blocks: we define these blocks as a Residual Block followed by a modified self-attention operation that is modulated by the artifact attention outputs. Importantly, this module receives feature maps generated by a global convolutional stem of three 3×3 convolution layers with strides 2, 1 and 1. The layers use the Mish [26] activation function and Batch Normalization layers. A 3×3 max pooling layer with stride 2 follows the stem, and a sequence of L Attention Blocks followed by a global average pooling and a sigmoid activation composes the rest of the network. We consider two types of Residual blocks in our architecture: the common Residual block of two 3×3 convolution layers, and the bottleneck block of three convolution layers with kernel size of 1×1, 3×3 and 1×1. We analyzed Attention block sequence sizes of $L = 18$ and $L = 34$ in our model. 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.

Self-attention with artifact heatmaps. We define our self-attention layers in a similar manner to prior self-attention work [6, 45], but extend the traditional construction to incorporate modulation from the artifact attention heatmaps. Specifically, our self attention layer computes an affinity matrix between keys and queries where the keys are re-weighted by the artifact attention map. The rationale here is that if human attention maps point to valuable areas in the frame, we want to maximize query affinity in those areas and inhibit it in others. Given a feature map $\mathbf{x_i}$ over one frame, and an artifact attention map A over that frame,



(a) Classifier module (b) Attention Block

Figure 5. **The classifier module and its attention blocks**. (a) our classifier takes the feature maps outputed by our convolutional stem, passes them through attention blocks modulated by human heatmaps, and computes logit scores for each frame before classifying the video. The consensus operation is instantiated as an average of the logits followed by thresholding to determine the output label. (b) Our attention block: a traditional residual block is followed by key, query and value matrices following the self-attention framework. The key-query product is modulated by our human heatmaps. $\times$ represents matrix multiplication and $\cdot$ represents element-wise multiplication.



(a) Artifact Attention and Caricature modules (b) Frame Distortion Block

Figure 6. **Our artifact attention and caricature modules**. (a) The artifact attention and caricature modules are intertwined: artifact attention operates with an encoder-decoder architecture to generate artifact heatmaps. Those heatmaps are supervised with the human heatmaps collected through our annotation interface. The caricature module receives both the heatmaps and the internal codes e, distorts those codes according to the artifact attention heatmaps, and generates caricatures by reconstructing the video from the distorted codes. (b) Our frame distortion block computes the difference between codes e_i and e_{i+1}, re-weights it according to the artifact attention maps, and then amplifies it by a factor of α before summing it back to e_i to generate distorted code $\hat{e}_{i+1}$.

the module learns to generate an affinity matrix $\mathbf{a_i}$:

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

The softmaxed key-query affinity tensor is then matrix-multiplied with the values $V = W_V x_i$ to generate the output residual r. That residual is then scaled by γ and added to input x_i to yield the output feature map y_i:

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

$\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V}$ are learned weight matrices of shape $\mathbf{R}^{\bar{C}\times C}$, $\mathbf{R}^{\bar{C}\times C}$ and $\mathbf{R}^{C\times C}$ respectively, with $\bar{C} = C/4$. γ is a learnable scalar which controls the impact of the learned attention feature vis-a-vis the original feature maps x_i.

4.2. Artifact Attention module

This 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. Our module is based on the encoder-decoder architectural pararadigm, and consists of an Xception-based encoder [10] and an 6-block resnet-based decoder, where upsampling and convolutions are layered between each block (Figure 6). The encoder produces codes e that retain spatial location information, and the decoder utilizes those compressed feature maps to generate output heatmaps. We use the human data collected through our interface to supervise the generation of these heatmaps directly. This happens though three losses: the Pearson Correlation Coefficient, KL-Divergence and L-1 loss. Our modules are trained

jointly, and the full loss is the combination of the cross-entropy loss in the classification pathway with the heatmap losses mentioned above.

4.3. Caricature module

Our caricature module is inspired by the motion magnification network proposed in [30]. The encodings e generated by the artifact attention module are used to amplify differences between the representations of consecutive frames. The artifact attention maps guide this module: element-wise multiplication with the artifact attention map of frame x_i is utilized to re-weight the difference in codes for each pair of frames x_i and x_{i+1} The distorted code $\hat{e}_i$ of frame x_i is computed as

$$\hat{\mathbf{e}}_\mathbf{i} + \mathbf{1} = \mathbf{e}_\mathbf{i} + \alpha(\mathbf{e}_\mathbf{i} - \mathbf{e}_{\mathbf{i+1}}) \odot \mathbf{A}, \tag{3}$$

where α is a user-defined distortion factor that controls the strength of the resulting caricature.

5. Experiments

To evaluate our deepfake detection framework, we (a) compare its detection performance to existing baselines and competing alternatives, and (b) evaluate the effectiveness of the caricatures. Performance is evaluated as classification accuracy in a real/fake classification task. We additionally ask whether a classifier trained on one dataset generalizes to others. We test human detection accuracy on the generated caricatures through a crowdsourced experiment, and we evaluate if deepfakes are indeed more detectable by humans when transformed into caricatures.

5.1. Benchmarks

We evaluate models on four benchmarks:

The Deepfake Detection Challenge Dataset (DFDC) [13] 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 [32] contains more than 1.8M images generated from 5000 YouTube videos. It features four different facial manipulation methods, which are Deepfake [1], Face2Face [39], FaceSwap [2] and NeuralTextures [38], 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 [21] dataset provides 5639 synthesized fake videos of celebrities generated from 590 real videos. Based on a tandem auto-encoder architecture, a video synthesis method that puts special emphasis on reducing 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 CariNet to the following baselines: the XceptionNet face detection model [32] and two variations of the ResNet model, which operate on single image frames or whole videos respectively.

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 [32].

Frame-based ResNet: We use ResNet [17] 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 [9, 16, 41]. As a baseline of 3D CNN, We use a 3D ResNet model where the 3D kernels are inflated from ResNet18 pretrained on ImageNet [12]. We also include a ResNet3D pretrained on the Kinetics dataset [8] as another baseline.

5.3. Implementation details

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 [34], 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. Our models were implemented in PyTorch, and our frame-based models were pretrained on ImageNet. All models were optimized with Rectified Adam [23] with LookAhead [46], 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. The 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. Detection performance

Table 1 shows the binary classification accuracy of the baseline models on the testing datasets as well as the overall performance, which corresponds to training on all training sets and evaluating on the combination of their test sets. We tested two versions of CariNet, changing the number of Attention Blocks: CariNet18 (18 Attention Blocks) and CariNet34 (34 Attention Blocks). CariNet34 achieves top performance on each dataset. Interestingly, the clip-based models were mostly outperformed by the majority of frame-based models. The ResNet3D model pretrained on Kinetics achieved high accuracy on the FaceForensics++ dataset, but lagged behind in other benchmarks. While the large-scale Kinetics dataset improves learning in videos, further analysis is needed to determine the reason behind its underperformance. 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.

5.5. Generalization to other datasets

Deepfake classifiers usually fail to generalize well to datasets built with different techniques from what the classifier was trained on. To test if our framework generalizes across deepfake generation techniques, we trained our module on one benchmark dataset and tested on the remaining three datasets. We tested our CariNet34 and compared it to two baselines (ResNet34 and XceptionNet, as above). Overall, we find that our setup improves the generalization abilities of baseline classifiers (Table 2).

5.6. Human detection results on deepfake caricatures

We ran 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 self-attention maps learned internally by our classifier (no artifact attention module), and 3) **Uniform caricature**: caricature videos where we uniformly distort the frames, without any spatial targeting induced by our human maps. We compare this to caricatures generated by our full CariNet pipeline.

Table 3 presents the results according to signal detection theory, a typical setup for human experiments.

We observe that our generated caricatures on deepfake videos significantly increase human detection of deepfakes, while maintaining the False Alarm rate relatively low. Our CariNet 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 $p < .0001$). Our 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 the participants do not have a biased response, 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, and have a tendency to respond *real* more often. The results also emphasize the importance of following our 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.

We additionally confirm qualitatively that our system generates large, visible distortions on fake videos while leaving real videos virtually untouched. As caricatures are at their best in video form, we point the reader to our video gallery in the supplemental for examples.

6. Ethical Impact and Limitations

Broadly, the goal of this work is to increase the trustworthiness of video-based media, by developing an automated procedure that unmasks doctored information on its initial presentation. There are several key features to our caricature-based approach. First, it integrates a signal about the trustworthiness of the presented information with the source of the information itself, by directly distorting the video. Second, it allows observers to detect for themselves that a video is doctored, rather than relying on third-party information about the video. Together, these feature may change how the video is encoded, which may prevent the misinformation from being absorbed and recalled. Additionally, caricatures are known to be more memorable [7, 25, 36, 42], which help people remember that a particular video was fake.

	Model	DFDC	FF++	Celeb	YouTube	Overall
Frame-based	XceptionNet	58.0	98.93	67.57	91.0	68.63
	ResNet18	93.25	82.86	96.91	100	93.26
	ResNet34	96.00	90.00	96.72	100	95.68
Clip-based	ResNet3D (ImageNet)	90.73	78.57	90.73	100	90.01
	ResNet3D (Kinetics)	88.75	96.07	92.66	92.85	92.58
CariNet (ours)	CariNet18	94.25	96.29	95.09	100	97.77
	CariNet34	**98.11**	**99.12**	**98.69**	**100.0**	**98.21**

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-CariNet corresponds to a version of the network where the magnification weights are frozen.

Model	FF++	Celeb	YouTube
RN34 (Baseline)	57.50	82.24	76.12
Xception (Baseline)	65.41	82.44	77.60
CariNet34 (Ours)	**69.07**	**84.31**	**80.13**

(a) DFDC

Model	DFDC	Celeb	YouTube
RN34 (Baseline)	52.25	70.84	64.28
Xception (Baseline)	51.03	67.43	**66.48**
CariNet34 (Ours)	**54.19**	**72.34**	66.21

(b) FF++

Model	DFDC	FF++	YouTube
RN34 (Baseline)	56.75	8.57	53.57
Xception (Baseline)	56.13	34.51	69.98
CariNet34 (Ours)	**58.02**	**39.64**	**82.14**

(c) Celeb

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).

Our human-centric approach may also be more robust to the constant and rapid improvement of deepfake quality than other systems - as long as deepfakes are creating frame-by-frame distortions which are perceptible to human annotators, our approach will be able to identify and amplify them. However, it will be important to continue to train our system on the latest datasets to ensure continued effectiveness across a wide variety of deepfake generation techniques.

As with many misinformation detection systems, there is risk that our networks could be reverse engineered to produce higher quality deepfakes. However, a system which allows humans to directly detect if a video is doctored will empower more individuals to assess for themselves whether to trust and engage with a given video. Aggregated over the many millions of people who watch videos online everyday, we believe that the benefits of such a system outweigh the risks.

7. Conclusion

We introduced CariNet, a network that can both detect deepfakes and distort them through the use of human artifact maps. Our network achieves state-of-the-art peformance on four different datasets, and the generated Deepfake Caricatures prove to be helpful to humans when trying to detect fakes, while avoiding distortions on real videos. It is important to note that 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 believe that our novel technique is a first of its kind approach, and that it has applications beyond simple deepfake detection. Our work establishes the feasibility of building relevant image manipulations that aid human perception.

References

[1] Deepfake. https://github.com/deepfakes/faceswap, 2020. 6

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap/, 2020. 6

[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

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

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

[6] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019. 4

[7] 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

[8] 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

[9] 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

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

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

[12] 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

[13] 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, 6

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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

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

[21] 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. 6

[22] 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

[23] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 6

[24] 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

[25] Robert Mauro and Michael Kubovy. Caricature and face recognition. *Memory & Cognition*, 20(4):433–440, 1992. 3, 7

[26] Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 2019. 4

[27] 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

[28] 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

[29] 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

[30] 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, 6

[31] 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

[32] 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, 6

[33] 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

[34] 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

[35] 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

[36] 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, 7

[37] 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

[38] 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. 6

[39] 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. 6

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

[41] 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

[42] Barbara Tversky and Daphna Baratz. Memory for faces: Are caricatures better than photographs? *Memory & cognition*, 13(1):45–49, 1985. 3, 7

[43] 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

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

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

[46] Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. *arXiv preprint arXiv:1907.08610*, 2019. 6