Video Manipulations Beyond Faces: A Dataset with Human-Machine Analysis

Trisha Mittal *
University of Maryland
trisha@umd.edu

Ritwik Sinha
Adobe Research, San Jose
risinha@adobe.com

Viswanathan Swaminathan
Adobe Research, San Jose
vishy@adobe.com

John Collomosse
Adobe Research, San Jose
collomos@adobe.com

Dinesh Manocha
University of Maryland
dmanocha@umd.edu

Abstract

As tools for content editing mature, and artificial intelligence (AI) based algorithms for synthesizing media grow, the presence of manipulated content across online media is increasing. This phenomenon causes the spread of misinformation, creating a greater need to distinguish between "real" and "manipulated" content. To this end, we present VIDEOSHAM, *a dataset consisting of* 826 *videos (*413 *real and* 413 *manipulated). Many of the existing deepfake datasets focus exclusively on two types of facial manipulations—swapping with a different subject's face or altering the existing face.* VIDEOSHAM, *on the other hand, contains more diverse, context-rich, and human-centric, high-resolution videos manipulated using a combination of* 6 *different spatial and temporal attacks. Our analysis shows that state-of-the-art manipulation detection algorithms only work for a few specific attacks and do not scale well on* VIDEOSHAM. *We performed a user study on Amazon Mechanical Turk with* 1200 *participants to understand if they can differentiate between the real and manipulated videos in* VIDEOSHAM. *Finally, we dig deeper into the strengths and weaknesses of performances by humans and SOTA-algorithms to identify gaps that need to be filled with better AI algorithms. We present the dataset here*[1].

1. Introduction

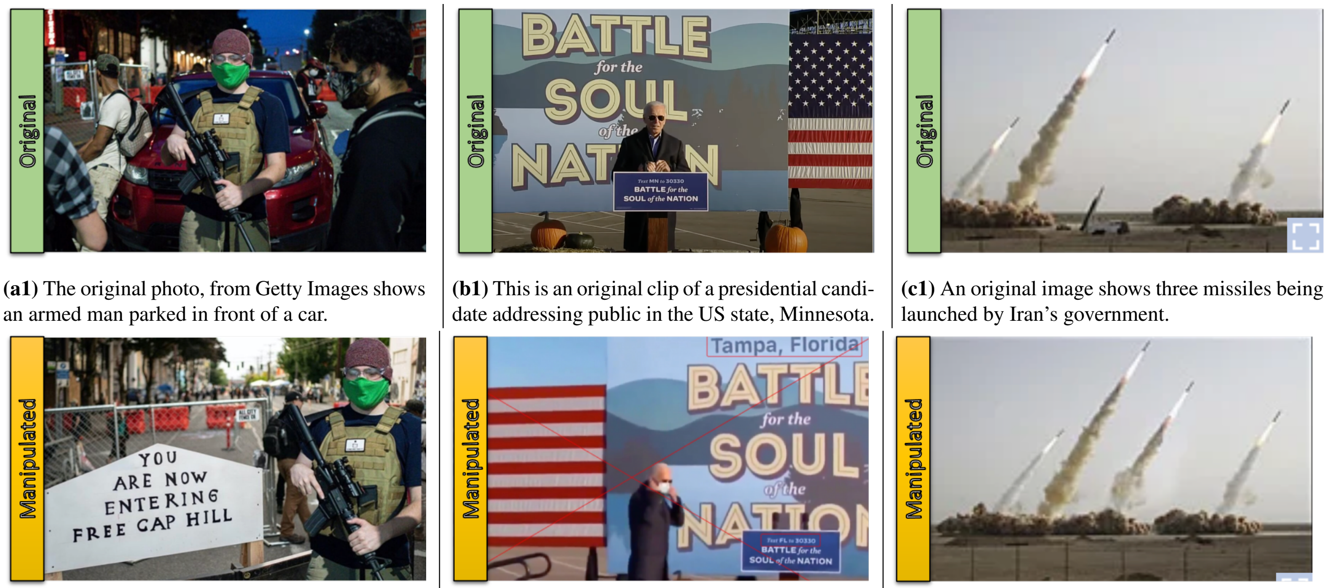
The proliferation of accessible video editing software and artificial intelligence (AI) tools has led to an increase in manipulated video content [23, 20]. While digital manipulation is commonplace in the creative process, in some cases video manipulation has a malicious intent. Social media often amplifies such false information through the circulation of manipulated videos [7, 14]. A recent survey by Pew Research Center showed that exposure to such false information is of widespread concern [54]. Therefore, there has been a significant increase in cases of misinformation, fraud and cybercrimes in the last decade. Such video manipulations pose a great threat to politics and can manipulate elections [62, 5], alter political narratives, weaken the public's trust in a country's leadership, and an increase hatred among various social groups. Another common occurrence is corporate frauds and scams where people use altered audio to impersonate other people to extort cash and other resources. Lastly, many video manipulations often result in numerous cybercrimes [18, 55, 9]. To further illustrate our motivations in this work, we depict such instances of video manipulations in Figure 1.

This leads to an important question—how do we detect manipulated content? The current arsenal of techniques involve the use of AI which in turn requires tremendous amounts of data. In the past decade alone, there has been a surge in the number of benchmark deepfake datasets [27, 49, 13] which manipulate the facial features of subjects in images and videos. We summarize recent deepfake datasets in Table 1.

But facial manipulations represent *only a fraction* of all manipulated content circulated on social media. For example, modifications also include changing the background context (Figure 1b), text and audio (Figure 1c) in media, aesthetic edits, adding/removing entities (Figure 1a), and temporal edits (Figure 1d). These manipulations can be performed in a matter of clicks due to the availability of state of the art video editing tools like Adobe AfterEffects^{TM}, Adobe Lightroom^{TM}, Filmora, GIMP, and many others. To our knowledge, no benchmark video dataset exists that extends beyond deepfake-only facial manipulations to include the vast range of manipulations described above.

[*] Work done as an intern at Adobe Research.

[1] VIDEOSHAM dataset link.


Original
BATTLE for the SOUL of the NATION
Original
Original
Manipulated
YOU ARE NOW ENTERING FREE CAP HILL
Manipulated
Tampa, Florida
BATTLE for the SOUL of the NATION
Manipulated

(a1) The original photo, from Getty Images shows an armed man parked in front of a car.

(b1) This is an original clip of a presidential candidate addressing public in the US state, Minnesota.

(c1) An original image shows three missiles being launched by Iran's government.

(a2) The photo above was altered by digitally placing the armed man in front of a peaceful protest, insinuating violence.

(b2) The clip above is altered by changing the location and the signs on the podium to a different US state, Florida.

(c2) In an altered image released on Iran's Revolutionary Guards website, claimed that 4 missiles were launched simultaneously.

Figure 1: **Spatial manipulations: (a)** [10], **(b)** [42], and **(c)** [44] are examples of videos on social media spatially manipulated with the intent to mislead audiences.

Faces	**Datasets**	**Release Date**	**# Videos**		**Source**		**Attacks**	**Human Density**	**Context**	**Modality**	
			Real	Fake	Original	Manipulated				Visual	Audio
Only	UADFV [66]	Nov-18	49	49	YouTube	Deep Learning	3	1	✗	✓	✗
	DF-TIMIT [27]	Dec-18	640	320	VidTIMIT [52]	Deep Learning	3, 4	1	✗	✓	✓
	FaceForensics++ [49]	Jan-19	1000	4000	YouTube	Deep Learning	3, 4	1	✗	✓	✗
	DFD [1]	Sep-19	0	3000	YouTube	Deep Learning	3	1	✗	✓	✗
	CelebDF [33]	Nov-19	5907	5639	YouTube	Deep Learning	3	1	✗	✓	✗
	DFDC [13]	Oct-21	23654	104, 500	Actors	Unknown	3	1	✗	✓	✗
	DeeperForensics 1.0 [21]	Jan-21	50, 000	10, 000	Actors	Deep Learning	3	1	✗	✓	✗
	WildDeepFake [72]	Jan-21	3, 805	3, 509	Internet	Internet	3, 4, 5	1	✗	✓	✗
	KoDF [30]	Aug-21	62, 166	175, 776	Actors	Deep Learning	3, 4, 5	1	✗	✓	✓
	FakeAVCeleb [22]	Sep-21	490+	20, 000+	VoxCeleb2 [11]	Deep Learning	3, 4	1	✗	✓	✓
	ForgeryNet [20]	July-21	91, 630	121, 617	Multiple	Deep Learning	3, 4	1	✗	✓	✓
	SR-DF [60]	Apr-21	1, 000	4, 000	YouTube	Deep Learning	3, 4	1	✗	✓	✓
	Khelifi et al. [23]	Jan-19	200	200	Multiple	User Generated	6, 7	1	✗	✓	✗
Beyond	MTVFD [4]	2016	30	30	YouTube	User Generated	1, 2	≤ 1	✓	✗	✗
	Liao et al [34]	2013	10	8	Multiple	User Generated	1	≤ 1	✓	✓	✗
	Su et al [56]	2015	7	7	SONY DSCP10	User Generated	1	≤ 1	✓	✓	✗
	Ours	Nov-21	413	413	Online Videos	User Generated		upto 40	✓	✓	✓

Table 1: **Characteristics of Video Manipulation Datasets:** We compare VIDEOSHAM with state-of-the-art video manipulation datasets.

	S.No.	**Attack**	**Method/Software**	**Description**
Spatial	1	Copy-Move and Splicing	Adobe Photoshop^{TM}, AfterEffects^{TM}	Select and copy region within same video and paste this somewhere else within the same video or different video
	2	Retouching/Lighting	Adobe Lightroom ^{TM}	Brightness increase/decrease, Contrast Increase/Decrease, Median Filter
	3	Face Swapping (FS)	FakeApp, FaceSwap [28] FaceShifter [31], FSGAN [43], DeepFaceLab [45]	Transferring a face from source to target image/video
	4	Face Re-enactment (FR)	Neural Textures [57], First-Order-Motion [53] Face2Face [58], IcFace [59], FSGAN [43],	Using facial movements and expression deformations of a face to guide the motions and deformations of another face
	5	Audio-driven FR (AFR)	Wav2Lip [46], APB2FACE [69], ATFHP [68]	Reenacting faces driven by a given audio signal to sync with lip movement
Temporal	6	Temporal	Adobe Lightroom ^{TM}	Frame Dropping, Frame Insertion, Shifting in time, Frame Swapping
Geometric	7	Geometric	Adobe Lightroom ^{TM}	Cropping, Resizing, Rotation, Shifting

Table 2: **Attacks:** We summarize the various attacks that have explored in prior literature for manipulating images and videos.

Main Contributions

We release a new manipulated high-resolution video dataset called VideoSham (Figure 6). VideoSham offers the following benefits over existing manipulated video datasets:

1. Beyond Faces (Deepfakes): The videos in VideoSham are manipulated using six spatial and temporal attacks (See Table 2) manipulating videos at the scene level targeting, not just faces, but also the background context, text and audio, aesthetic edits, adding/removing entities, and temporal edits (See Figure 2).

2. Beyond Images: Although there exist image manipulation datasets that go beyond faces, they cannot be used to detect video manipulations, which require dedicated video datasets. The latter, however, are hard to create due to the manual labor involved. In this work, we go beyond images to release the first video manipulation dataset containing beyond-face manipulations.

VideoSham consists of 413 real-world videos and their corresponding manipulated versions (total 826 videos). The videos have diverse scene backgrounds, are context-rich, and contain up to 9 subjects on average. VideoSham is the largest dataset containing manipulated videos generated by professional video editors with varied attacks. A user study conducted on Amazon Mechanical Turk (AMT) to understand the kind of attack methods that mislead humans the most. In addition, we analyze the performance of existing state of the art deepfake detection algorithms and video forensics algorithms on VideoSham. We find that these techniques are less than 50% effective in distinguishing between a real and a manipulated video.

We elaborate more about VideoSham in Section 3. In Section 4 we present our findings from the user study and evaluation of detection models. And, finally in Section 5, we discuss some promising ideas to help these attacks.

2. Related Work

In this section, we discuss previous works in detection of manipulated and deceptive media content. To begin with, we first discuss the video manipulation techniques used to create such fake videos in Section 2.1. Then in Section 2.2, we summarize various datasets and benchmarks for video manipulations. We also survey different techniques used for detecting deepfake videos in Section 2.3 and generic video forensic methods in Section 2.4.

2.1. Video Manipulation Techniques/Attacks

Manipulation techniques, or attacks, are broadly categorized as spatial [6], temporal [23], and geometric [23] in the literature (see Table 2). Basic examples of spatial attacks include copy-move and image/video splicing which correspond to spatially or temporally shifting an object to a different location in the same video or a different video, respectively. Retouching, another common attack, involves aesthetic edits like adjusting brightness, contrast, and other parameters of digital content. More recently, people have used AI to alter facial features to create deepfake videos. AI-based techniques are comprised of two major attack approaches, Face Swapping [43, 45] and Face Reenactment [58, 57, 53]. Face Swapping switches the subject's face with the face of another person and Face Reenactment alters the subject's facial expressions. Temporal attacks involve swapping, duplicating, inserting, and deleting frames of video, giving the impression that the video has been sped up or slowed down. Finally, geometric attacks include operations like cropping and rotations.

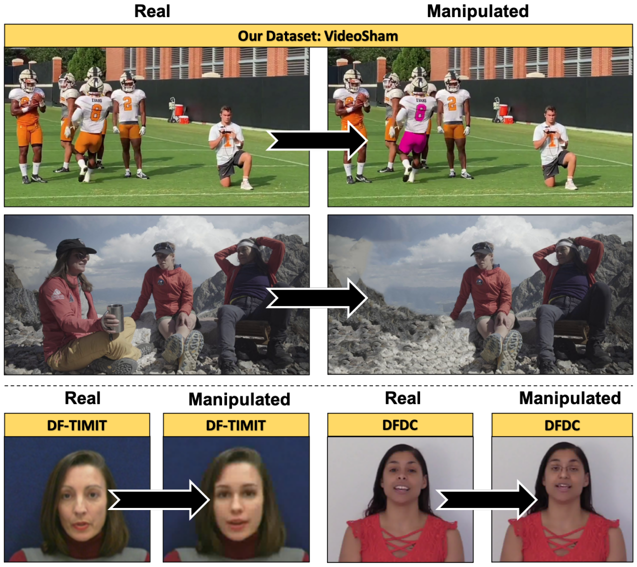


Figure 2: **VideoSham:** *(top)* VideoSham consists of diverse, context-rich, and human-centric manipulated videos by professional video editors via 6 spatial and temporal attacks (*e.g.* jersey color change and person removal). *(bottom)* In contrast, deepfake datasets (DF-TIMIT and DFDC) only consist of facial manipulations individual subjects from a close-up angle.

2.2. Video Manipulation Datasets

Creating benchmarks of video manipulations is a challenging task as this may require per-frame manipulations. Some of the datasets (like Khelifi et al. [23], MTVFD [4], Liao et al. [34], Su et al. [56], Media Forensics Challenge [15]) are very small in volume containing $7 - 200$ videos each, these datasets are also **not publicly available**. Most of these videos have 0 or 1 subjects present in the frame with very little background context. More recently, AI-synthesized attacks like *face swapping*, *face re-enactment*, and *audio-driven face re-enactment* have led to the creation of datasets like UADFV [66], FaceForensics++ [49], DeeperForensics1.0 [21], WildDeepFake [72]. Because these datasets are generated using learning methods; some of these datasets have upto 100k videos. However all of these datasets have strictly 1 subject per video with the face being predominant part of the frame with no background context at all. Many datasets are missing audio except DFDC [13], DF-TIMIT [27], KoDF [30], FakeAVCeleb [22], ForgeryNet [20] and SR-DF [60].

2.3. Deepfake Detection Methods

The goal of the deepfake detection approaches is to algorithmically distinguish fake videos from real videos. A large portion of these methods are focused on detecting visual artifacts especially on the finer regions of the face (like eyes, mouth and teeth [39]). Some approaches specifically focus on abnormalities like inconsistent head pose orientations [67], asynchronous lip movement and speech [17] and unnatural eye blinking [32]. Prior work have also observed and exploited the fact that temporal coherence is not enforced effectively in the synthesis process of deepfakes [51, 16] and exploit this in detection methods. More recently, interesting affective computing approaches that focus on correlated emotion signals from audio-visual cues [40, 3], and detecting signals like heart rate and breathing rate [47] from the videos have also been proposed. However, it is clear that due to the nature of the datasets (single-person, face-centered videos), these approaches focus only on facial cues and audio cues.

2.4. Video Forensic Methods

Developments in video forensics literature focus on two specific attacks; Copy-Move and Splicing (Row 1 in Table 2) and Temporal attacks (Row 6 in Table 2). Most conventional copy-move forgery detection methods mainly consist of three components [12]: (1) feature extraction, (2) matching, and (3) post-processing. A variety of features have been explored, e.g., DCT (Discrete Cosine Transform) [38], DWT (Discrete Wavelet Transform) and KPCA (Kernel Principal Component Analysis) [8], Zernike moments [50]. Consequently, some end-to-end deep learning based copy-move forgery detection methods were proposed [63, 64, 32]. However these efforts are limited to images. Another interesting development, still in naive stages is deep learning methods to detect inpainting in videos [71]. Some of the methods in detecting temporal attacks (also called as intra-frame manipulations) use the consistency of velocity field [65] and optical flow [61]. These methods can recognize frame insertion and frame deletion attacks. Similarly, Zhao et al. [70] use inter-frame similarity analysis to detect frame duplications in the videos. Finally, Long et al. [37] propose a coarse-to-fine framework based on deep Convolutional Neural Networks (CNN) to detect potential frame duplications.

3. Our Dataset- VIDEOSHAM

In this section, we present details on the dataset creation process (Section 3.1) and discuss some of the salient features and characteristics of VIDEOSHAM (Section **??**).

3.1. Creation and Annotation Process

3.1.1 Source Videos

We have a total of 836 videos comprising of 413 original videos and 413 manipulated versions, each corresponding to one of the original videos. We obtain our source videos from an online video website (vimeo [2]) and only include videos attributed with a CC-BY (Creative Commons) license. In addition, we avoid videos with brands, children, objectionable content, TV show/movie clips and videos with copyrighted music. We trim these original videos to a specific length (upto $5-30$ seconds) before we perform any manipulation attack.

3.1.2 Manipulation Attacks

We employ a total of 6 manipulation attacks for creating our dataset. As per prior literature, we also categorize these attacks into spatial and temporal attacks[3]. We visually show the distribution of the attacks in Figure 3a (attacks outlined in blue are spatial attacks, outlined in pink are temporal attacks). We describe each of the attack below.

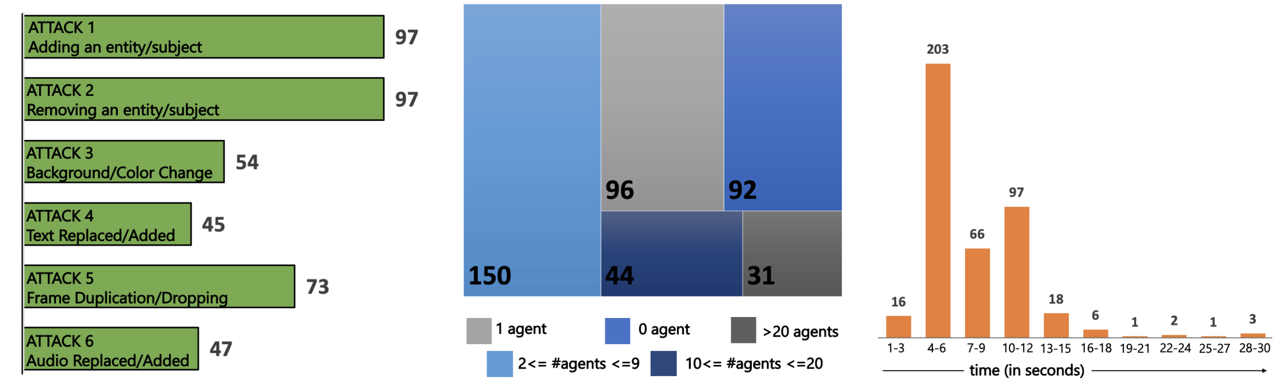
- ATTACK 1 (Adding an entity/subject): In this attack we select an entity or a subject from some other sources and place them in the current video. This attack is somewhat similar to copy-move attack.

- ATTACK 2 (Removing an entity/subject): In this attack, we basically select an entity or a subject in the video and remove it from all the frames and fill in the gap with background settings. To do this, we used content-aware fill in Adobe AfterEffects$^{\text{TM}}$ and some deep learning methods for generating masks [19] and performing video inpainting [24, 25].

- ATTACK 3 (Background/Color Change): We focus on a particular aspect of the video, and change the background of the video, or color of a small entity in the video.

[2] www.vimeo.com.

[3] We do not use geometric attacks, as they have been shown to be easily detected.



(a) **Attack distribution:** Distribution of videos that are attacked with different manipulation techniques. Attacks $1-4$ are spatial attacks, and Attacks $5-6$ are temporal attacks.

(b) **Density distribution:** Distribution of videos according to number of persons present in each video. This is considerably high w.r.t. the existing datasets.

(c) **Duration distribution:** Distribution of videos according to duration or length of each video (in seconds). The average length of our videos is 8 seconds.

Figure 3: **Dataset statistics:** We visually present various statistics for VIDEOSHAMfor better insights.

– ATTACK 4 (Text Replaced/Added): We perform edits like adding some text in the video or removing or replacing already existing text in the video.

– ATTACK 5 (Frames Duplication / Removal/ Dropping): This attack is specifically to render the video temporally inconsistent. We choose to perform one of these manipulations, randomly duplicating frames, removing or dropping frames in the video. This also includes slowing down a video.

– ATTACK 6 (Audio Replaced): Audio modality is a very important aspect for videos. To manipulate this, we replace the existing audio with some other audio.

We visually depict the 4 spatial attacks (ATTACK 1, ATTACK 2, ATTACK 3, and ATTACK 4) in Figure 6.

3.1.3 Manipulated Videos

We worked with 3 professional video editors hired on Upwork [4]. The editors were shortlisted based on their experience and were well-versed with Adobe AfterEffects$^{\text{TM}}$, the software used for creating these edits. Each editor was assigned tasks, i.e. source videos, start and end timestamp to be edited and a one-line description of the manipulation to be performed. We provide all videos and the attacks performed for every video.

Dataset Analysis: In Figure 3a, we present the distribution of attacks for the 413 videos, each lasting $1-31$ seconds. The average length of videos in our dataset is around 8 seconds long. We also run an object detection model [5] to count the number of people/agents in every video (Figure 3b). More than 80% of the videos in our dataset contains at least one subject.

4. Experiments and Results

We elaborate on three experiments we perform to highlight the importance, novelty and usecase of VIDEOSHAM. To begin with, we present the analysis of how well humans fair in detecting these attacks in Section 4.1, followed by analysis of the performance of state-of-the-art deepfake detection methods and video forensic techniques in Section 4.2. Finally in Section 4.3, we present some ideas and preliminary results for using interdisciplinary ideas for detecting such attacks.

4.1. Expt 1: How Well Do Humans Perform?

Setup: We first shortlist 60 videos from VIDEOSHAM. Out of these, 30 videos are real and the remaining 30 are manipulated (5 videos per attack). We recruit human participants from Amazon Mechanical Turk (AMT) and show each video to 20 participants. The participants are requested to watch the full video; followed by two questions. In the first question, the participants are asked to respond to the following prompt in either a yes or no - “Do you believe this video has been manipulated/edited to misrepresent facts?”. And, in the second question we ask them to explain in a sentence what they felt was manipulated with the following prompt- “If you answered YES above, what region or aspect of this video, do you believe is manipulated.”. Note that participants are not informed whether videos are manipulated or not. They are also not informed about the set of attacks. We show the setup in Figure 4 which was used to collect a total of 1200 responses from AMT participants.

[4] www.upwork.com.

[5] https://github.com/roboflow-ai.

(A) 1200 *responses* (20 participants × (30 real + 30 manipulated videos))

GT		#Resp (Total)	Rep Real	Rep Fake	p_1	p_2	$p_1 - p_2$	$p-$ value	CI(l)	CI(u)
Real		600	454	146	0.757		0	–	–	–
Manipulated Attack	1	100	87	13	0.757	0.87	−0.113	0.991	−0.182	1
	2	100	79	21	0.757	0.79	−0.033	0.725	−0.112	1
	3	100	74	26	0.757	0.74	0.016	0.408	−0.066	1
	4	100	69	31	0.757	0.69	0.066	0.097	−0.020	1
	5	100	75	25	0.757	0.75	0.006	0.493	−0.076	1
	6	100	55	45	0.757	0.55	0.207	$<$ **0.001**	0.114	1
		600	439	161	0.757	0.732	0.0250	0.177	−0.018	1

Table 3: **Human Performance:** We observe that participants are unable to detect ATTACK 3 (26%) and ATTACK 4 (31%). Videos manipulated using ATTACK 5 are relatively easier to detect (75%).

GT		#Vid.	**Predicted**									
			Deepfake Detection Methods						Video Forensics Techniques			
			Li et al [32]		MesoNet [2]		Mittal et al. [40]		Long et al. [37]		Liu et al. [36]	
			Predicted		**Predicted**		**Predicted**		**Predicted**		**Predicted**	
			Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake
Real		413	188	225	167	246	238	175	219	194	234	179
Manipulated	**Attack** 1	97	76	21	93	4	92	5	86	11	68	29
	2	97	63	34	84	13	66	31	84	13	46	51
	3	54	35	19	37	17	49	5	50	4	38	16
	4	45	32	13	34	11	42	3	39	6	34	11
	5	73	70	3	67	6	54	19	25	48	68	5
	6	47	45	2	44	3	31	16	38	9	41	6
		413	321	92	359	54	334	79	322	91	295	118

Table 4: **Machine Performance:** We evaluate 3 state-of-the-art deepfake detection methods and 2 video forensics techniques on VIDEOSHAM. It is apparent that these algorithms do not perform well on VIDEOSHAM, speaking to its complexity and diversity.

Study Analysis: We summarize the responses of the user study in Table 3. Both the real and manipulated videos receive 600 responses each. We observe that out of the 600 responses (corresponding to the 30 real videos), 342, i.e., 57% were correctly identified as real. Similarly, out of 600 responses for the manipulated videos, 389 were incorrectly identified as real, i.e., 35.2% of these responses correctly identified manipulated videos. Analyzing the responses by the type of attack, we observe that human participants are able to identify 45% of the videos manipulated using ATTACK 6. For the other attack types, the proportion of manipulated videos labeled as 'fake' ranges from 13-31%. Furthermore, we notice that human participants are able to more successfully identify manipulated videos that are modified using temporal attacks (ATTACK 5 and ATTACK 6) than spatial attacks (ATTACK 1− ATTACK 4). Moreover, we also received some number of responses from participants explaining their rationale behind reporting a manipulated video. From the responses received, there is no clear evidence that suggests that participants are able to identify the manipulated region/kind in case of spatial edits. But, they were somewhat able to correctly identify the manipulated edit in case of temporal attacks. This would imply that a subset of our selection of attacks are indiscernible to the human eye.

Statistical Tests: Next we consider statistical tests to see if humans are able to tell a real video from a manipulated video. We consider the following quantities for this test, define $p_1 = P(\text{declaring video real}|\text{real video})$. Also, let $p_2 = P(\text{declaring video real}|\text{manipulated video})$. If humans are able to tell real videos apart from manipulated videos, we expect p_1 to be larger that p_2. Hence, we test the one-sided statistical hypothesis:

$$H_0 : p_1 = p_2 \quad \text{against} \quad H_1 : p_1 > p_2.$$

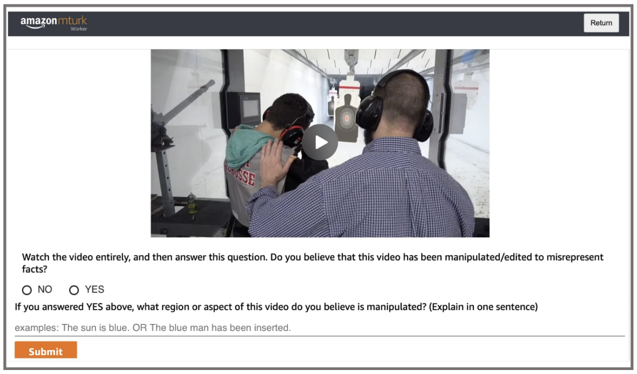
We test this hypothesis with the test statistic $(p_1 - p_2)$[6]. In Table 3 we present the difference of proportions as well as the one-sided $p-$value of the test for each attack type. The first thing to note is that when combining across all attack types (last row), we see that even though p_1 is slightly bigger than p_2, this difference is not statistically significant ($p-$value of 0.177). This suggests that our edits are not discernible to human evaluators. When we break it down by attack type, we observe that only for ATTACK 6 (audio replacement), humans are more likely to declare such edits as manipulations ($p-$value < 0.001). For ATTACK 4 (text replaced or added), there is weak statistical evidence of humans detecting this manipulation ($p-$value of 0.097). For all other attacks, there is no statistical evidence that humans can tell when a video has been manipulated using that strategy. It is particularly telling that when an entity/subject is added or removed (ATTACKs 1 and 2), more of our human subjects declare such manipulated videos as real than they declare unedited videos. This shows how modern editing tools can be used to manipulate videos in a way that humans have no way of telling such edits just by looking at the video. This observation establishes the need to build high quality video manipulation detection algorithms that can label manipulated videos at scale.

4.2. Expt 2: How Well Do Machines Perform?

To answer this question better, we evaluate state-of-the-art deepfake detection methods and video forensics techniques on VIDEOSHAM.

Deepfake Detection Methods: We evaluate Li et al. [32], XceptionNet [49] and Mittal et al. [40] on VIDEOSHAM. Deepfake videos generated using data-driven methods can only synthesize face images of a fixed size, and they must undergo an affine warping to match the configuration of the source's face. Due to resolution inconsistencies between warped face and background context, there are various artifacts on the synthesized faces. Li et al. [32] detects such artifacts by comparing the generated face areas and their surrounding regions with a dedicated Convolutional Neural Network (CNN) model. On the other hand, XceptionNet [49] is a transfer learning model which is also a CNN

[6] Given our sample size, we have a 86% statistical power of detecting a difference if the true values are $p_1 = 0.75$ and $p_2 = 0.74$.



Watch the video entirely, and then answer this question. Do you believe that this video has been manipulated/edited to misrepresent facts?

○ NO ○ YES

If you answered YES above, what region or aspect of this video do you believe is manipulated? (Explain in one sentence)

examples: The sun is blue. OR The blue man has been inserted.

Figure 4: **User Study Setup:** We present the Amazon Mechanical Turk setup used (Section 4.1).

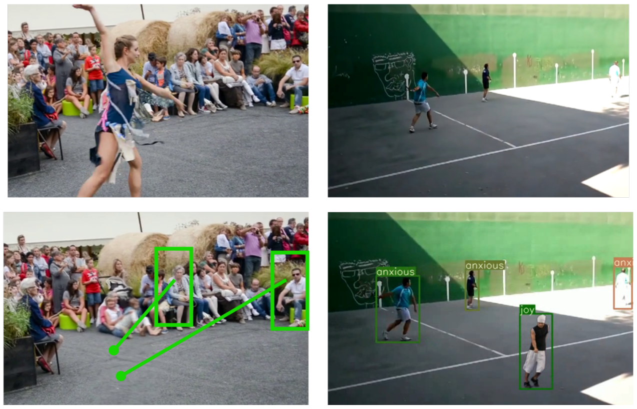


Figure 5: **Inter-Agent Dynamics and Multimodal Ideas for Detecting Manipulations:** We show the output of the automated techniques used to identify manipulated videos in VIDEOSHAM. *(Column 1)* In the first column, we remove the main subject from the foreground. We identify this as a manipulated image using a gaze tracking algorithm by noting that there is no object at the location of the crowds gaze direction. *(Column 2)*Here, we manipulate an image by inserting the man in black shirt. We use emotion recognition techniques to infer that this false subject has an affective state that is not in tune to those of the other players.

architecture, which was originally trained for the classical object detection task and later finetuned for deepfake detection on FaceForensics++ dataset. Finally, Mittal et al. propose an approach that simultaneously exploits the audio (speech) and video (face) modalities and also the perceived emotion features extracted from both the modalities to detect any falsification or alteration in the input video. They use the correlation between the modalities to detect a fake video.

GroundTruth	# videos	Reported Real	Reported Manipulated
Real	413	286	127
Attack 1	97	32	65
Attack 2	97	35	62

Table 5: **Quantitative Results (Expt 3):** For some preliminary analysis, we explore two ideas, *gaze* and *affect* of all agents involved. We observe that these two ideas in itself can effectively detect manipulations of the kind, ATTACK 1 and ATTACK 2.

Video Forensics Techniques: We evaluate Long et al. [37] and Liu et al. [36] on VIDEOSHAM. Both of these methods are state-of-the-art methods in video forensics literature. While, Long et al. [37] is specifically for detecting cases of frame duplications in a video, Liu et al. [36] specifically focus on detecting copy-move attacks. For all the methods, we use pretrained models and report the results when evaluated on VIDEOSHAMin Table 4.

Study Analysis: All the 5 shortlisted methods are less then 50% accurate on VIDEOSHAM. This is understandable, as all the deepfake methods (Li et al. [32], MesoNet [2], and Mittal et al. [40]) are trained specifically to look for manipulations in faces. Moreover, these method are not used to inferencing on videos with more or less than 1 person in the frame and with so much context information. Hence, we observe that these methods are only inferring based on artifacts caught near the face regions in the VIDEOSHAM videos. We also observe that, Mittal et al. specifically are able to detect some of the temporal manipulations well; which is because the method is trained to look for correlation between audio and visual modalities. Similarly, even the video forensics techniques are specifically performing well on attacks that they have been trained for, i.e. ATTACK 5 for Long et al. and ATTACK 1 and ATTACK 2 for Liu et al. [35]. ATTACK 3 (color change) and ATTACK 4 (text replacement) tend to remain hard to be detected by most of these methods.

4.3. Expt 3: Beyond DeepFake Detection and Video Forensic Techniques

One can observe from the experiments in the previous section, that all the methods are largely dependent on the visual artifacts. However, given the diversity of attacks used to manipulate videos, we hypothesize the use of inter-agent and multimodal analysis models for detecting such manipulations. We show preliminary results in Figure 5.

Strategy 1 (Gaze): To begin with, we believe that tracking gaze of subjects can be useful for detection experiments. Gaze following is a task in computer vision to identify objects and regions that the subject of interest is focusing on. The idea behind this strategy is to identify manipulated images by using gaze following to locate "absent" targets and/or "out-of-context" subjects in the video. To perform some preliminary analysis we deploy GazeFollow [48]. More specifically, for each frame, we begin by obtain the spatial coordinates of the subject's head's bounding box and pass this information as input to the gaze tracking algorithm, GazeFollow [48], which outputs the location of the subject's gaze. The final step in this strategy is to run an object detector to obtain a confidence score c_g corresponding to an object present at the gaze location. A low confidence score indicates a manipulated frame.

Strategy 2 (Affect): In this strategy we propose the use of affective cues. When we track and look for affective disparities in affective state of different subjects. Prior works in psychology [26] and empirical works [41] that subjects in social settings often share affective states. We use facial expressions, body postures and scene understanding to perceive the affective states of all subjects. We use the model EmotiCon [41] trained on EMOTIC dataset [29] to perceive these affective states and obtain an affective confidence score c_a. By empirically assigning a threshold, τ on the two confidence scores, we flag a video as manipulated. We observe that these two techniques help detect ATTACK 1 and ATTACK 2 significantly well. We add quantitative results for the same in Table 5. We show two qualitative results of these ideas in Figure 5.

Experiment 3 shows that, in addition to human assessment, specialized deepfake detection techniques, and video forensics, other approaches that are not intended for identifying manipulated videos can be used.

5. Conclusion and Future Directions

Our goal with the expt 1 (Section 4.1) and expt 2 (Section 4.2) was to understand how well humans can detect some of the manipulations that occur today circulated on social media. We also wanted to understand if the developments in the deepfake detection and video forensic literature match up to these manipulation attacks. Finally, through expt 3 (Section 4.3) we want to propagate the idea of using ideas beyond detection of visual artifacts for scalable models for video manipulation detection.

We conclude from expt 1 (Section 4.1) and expt 2 (Section 4.2) that both humans and machines (5 methods shortlisted) struggle to detect these manipulations successfully. We believe that these are attacks of concern, as they are going undetected even by human participants. Moreover, we emphasize that these manipulations play a big role in many real-world video manipulations (Figure 1).

More generally, we believe that computer vision algorithms perform almost comparable to humans in most of these ATTACKS. However, most methods are very attack-specific and do not generalize well to other attacks. Mostly every deepfake detection method fails to handle videos with more than 1 subject and hence have a very limited scope. Also, importantly most of the deepfake detection methods require huge amounts of training samples; and this is not a realistic assumption. It is important to build methods which can be less computationally intensive and at the same time are also able to generalize well. Similarly, methods in video forensics also are only able to handle very specific attacks. These are less dependent on data, but computationally expensive as they are more or less, inference based methods.

We believe following are some knowledge gaps and research agendas that can help the society combat the increasing problem of misinformation, frauds and cybercrimes occurring due to manipulated media content shared online.

1. There is a need to build detection models focused on more diverse attacks or video manipulations. Through VIDEOSHAM, we attempted to include some of the attacks that have not been studied before owing to a lack of a dataset. We hope this dataset can be a step towards achieving better detection models for all the 6 attacks.

2. Moreover it is important to increase the scope of detection ideas being used currently for detecting manipulations. Current methods are extremely focused on visual perception. Our goal through experiment 3 was to show through very preliminary analysis that ideas based on inter-agent dynamics and multimodal cues can be a promising literature source. Another promising idea, is to include domain knowledge in detecting manipulations; as humans we have some contextual information which the detection models severely suffer from.

3. Largely all existing methods require a significant amount of training data to train the models. But, with newer manipulations and attacks on videos, it will become impossible to keep up with detection models for the same. We need to reduce the dependence on training data build detection models that are as generalizable as possible to potential attacks.

6. Ethical Considerations

We note that our dataset sources videos from an online video website that are attributed with a CC-BY license, and we do not retain any metadata corresponding to the creators of the videos. In addition, we do not collect any personal information of the human participants in the subsequent user study conducted on AMT. We expect that our dataset is an effort towards mitigating and fighting against malicious manipulations of online digital content.

Acknowledgements

This research is partially supported by ARO Grant W911NF2110026.

References

[1] Contributing data to deepfake detection research, Sep 2019.

[2] 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.

[3] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2020.

[4] Omar Ismael Al-Sanjary, Ahmed Abdullah Ahmed, and Ghazali Sulong. Development of a video tampering dataset for forensic investigation. *Forensic science international*, 266:565–572, 2016.

[5] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14):eaay3539, 2020.

[6] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. A sift-based forensic method for copy–move attack detection and transformation recovery. *IEEE transactions on information forensics and security*, 6(3):1099–1110, 2011.

[7] Katie Anderson. Getting acquainted with social networks and apps: combating fake news on social media. *Library Hi Tech News*, 35, 07 2018.

[8] M Bashar, K Noda, N Ohnishi, and K Mori. Exploring duplicated regions in natural images. *IEEE Transactions on Image Processing*, 2010.

[9] Johnny Botha and Heloise Pieterse. Fake news and deepfakes: A dangerous threat for 21st century information security. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited*, page 57, 2020.

[10] Jim Brunner. Fox news runs digitally altered images in coverage of seattle's protests, capitol hill autonomous zone — the seattle times. Link, June 2020.

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[12] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, 10(11):2284–2297, 2015.

[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.

[14] Alvaro Figueira and Luciana Oliveira. The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121:817–825, 12 2017.

[15] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72. IEEE, 2019.

[16] 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.

[17] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection, 2021.

[18] Douglas Harris. Deepfakes: False pornography is here and the law cannot protect you. *Duke L. & Tech. Rev.*, 17:99, 2018.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[20] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4360–4369, 2021.

[21] Liming Jiang, Wayne Wu, Ren Li, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. *arXiv preprint arXiv:2001.03024*, 2020.

[22] Hasam Khalid, Shahroz Tariq, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[23] Fouad Khelifi and Ahmed Bouridane. Perceptual video hashing for content identification and authentication. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):50–67, 2017.

[24] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019.

[25] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Recurrent temporal aggregation framework for deep video inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1038–1052, 2020.

[26] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4:15–33, 2013.

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

[28] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017.

[29] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotic: Emotions in context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 61–69, 2017.

[30] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. *arXiv preprint arXiv:2103.10094*, 2021.

[31] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[32] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[33] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. 2019.

[34] Sheng-Yang Liao and Tian-Qiang Huang. Video copy-move forgery detection and localization based on tamura texture features. In *2013 6th international congress on image and signal processing (CISP)*, volume 2, pages 864–868. IEEE, 2013.

[35] Haomiao Liu, Ruiping Wang, S. Shan, and Xilin Chen. Deep supervised hashing for fast image retrieval. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2064–2072, 2016.

[36] Yaqi Liu, Chao Xia, Xiaobin Zhu, and Shengwei Xu. Two-stage copy-move forgery detection with self deep matching and proposal superglue. *IEEE Transactions on Image Processing*, 31:541–555, 2021.

[37] Chengjiang Long, Arslan Basharat, Anthony Hoogs, Priyanka Singh, Hany Farid, et al. A coarse-to-fine deep convolutional neural network framework for frame duplication detection and localization in forged videos. In *CVPR Workshops*, pages 1–10, 2019.

[38] Toqeer Mahmood, Tabassam Nawaz, Aun Irtaza, Rehan Ashraf, Mohsin Shah, and Muhammad Tariq Mahmood. Copy-move forgery detection technique for forensic analysis in digital images. *Mathematical Problems in Engineering*, 2016, 2016.

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

[40] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020.

[41] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14222–14231, 2020.

[42] AP News. False video of joe biden claims democrat candidate greeted wrong us state at a rally — euronews. Link, February 2020. (Accessed on 04/07/2022).

[43] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.

[44] NPR. Photo of iran's missile launch was manipulated : Npr. Link, July 2008. (Accessed on 04/07/2022).

[45] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[46] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020.

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

[48] Adriá Recasens Continente Recasens. *Where are they looking?* PhD thesis, Massachusetts Institute of Technology, 2016.

[49] 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.

[50] Seung-Jin Ryu, Matthias Kirchner, Min-Jeong Lee, and Heung-Kyu Lee. Rotation invariant localization of duplicated image regions based on zernike moments. *IEEE Transactions on Information Forensics and Security*, 8(8):1355–1370, 2013.

[51] 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, 2019.

[52] Conrad Sanderson. The vidtimit database. Technical report, IDIAP, 2002.

[53] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32:7137–7147, 2019.

[54] Laura Silver. Misinformation and fears about its impact are pervasive in 11 emerging economies, Aug 2020.

[55] Russell Spivak. " deepfakes": The newest way to commit one of the oldest crimes. *Geo. L. Tech. Rev.*, 3:339, 2018.

[56] Lichao Su, Tianqiang Huang, and Jianmei Yang. A video forgery detection algorithm based on compressive sensing. *Multimedia Tools and Applications*, 74(17):6641–6656, 2015.

[57] 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.

[58] 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.

[59] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3385–3394, 2020.

[60] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. M2tr: Multi-modal multi-scale transformers for deepfake detection. *arXiv preprint arXiv:2104.09770*, 2021.

[61] Qi Wang, Zhaohong Li, Zhenzhen Zhang, and Qinglong Ma. Video inter-frame forgery identification based on optical flow consistency. *Sensors & Transducers*, 166(3):229, 2014.

[62] Duncan J Watts, David M Rothschild, and Markus Mobius. Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

[63] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Busternet: Detecting copy-move image forgery with source/target localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018.

[64] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. long2019coarse. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1907–1915. IEEE, 2018.

[65] Yuxing Wu, Xinghao Jiang, Tanfeng Sun, and Wan Wang. Exposing video inter-frame forgery based on velocity field consistency. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2674–2678. IEEE, 2014.

[66] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[67] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[68] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020.

[69] Jiangning Zhang, Liang Liu, Zhucun Xue, and Yong Liu. Apb2face: audio-guided face reenactment with auxiliary pose and blink signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4402–4406. IEEE, 2020.

[70] Dong-Ning Zhao, Ren-Kui Wang, and Zhe-Ming Lu. Inter-frame passive-blind forgery detection for video shot based on similarity analysis. *Multimedia Tools and Applications*, 77(19):25389–25408, 2018.

[71] Peng Zhou, Ning Yu, Zuxuan Wu, Larry S Davis, Abhinav Shrivastava, and Ser-Nam Lim. Deep video inpainting detection. *arXiv preprint arXiv:2101.11080*, 2021.

[72] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2382–2390, 2020.