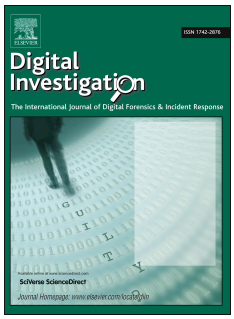
A review of digital video tampering: from simple editing to full synthesis.

JOHNSTON, P. and ELYAN, E.

2019

Accepted Manuscript



A review of digital video tampering: From simple editing to full synthesis

Pamela Johnston, Eyad Elyan

PII: S1742-2876(18)30414-6

DOI: https://doi.org/10.1016/j.diin.2019.03.006

Reference: DIIN 841

To appear in: *Digital Investigation*

Received Date: 16 November 2018

Revised Date: 24 January 2019

Accepted Date: 17 March 2019

Please cite this article as: Johnston P, Elyan E, A review of digital video tampering: From simple editing to full synthesis, *Digital Investigation* (2019), doi: https://doi.org/10.1016/j.diin.2019.03.006.

A Review of Digital Video Tampering: From Simple Editing to Full Synthesis

Pamela Johnston[a,*], Eyad Elyan[a]

[a]*Robert Gordon University, Garthdee House, Garthdee Road, Aberdeen, AB10 7QB, Scotland, UK*

Abstract

Video tampering methods have witnessed considerable progress in recent years. This is partly due to the rapid development of advanced deep learning methods, and also due to the large volume of video footage that is now in the public domain. Historically, convincing video tampering has been too labour intensive to achieve on a large scale. However, recent developments in deep learning-based methods have made it possible not only to produce convincing forged video but also to fully synthesize video content. Such advancements provide new means to improve visual content itself, but at the same time, they raise new challenges for state-of-the-art tampering detection methods. Video tampering detection has been an active field of research for some time, with periodic reviews of the subject. However, little attention has been paid to video tampering techniques themselves. This paper provides an objective and in-depth examination of current techniques related to digital video manipulation. We thoroughly examine their development, and show how current evaluation techniques provide opportunities for the advancement of video tampering detection. A critical and extensive review of photo-realistic video synthesis is provided with emphasis on deep learning-based methods. Existing tampered video datasets are also qualitatively reviewed and critically discussed. Finally, conclusions are drawn upon an exhaustive and thorough review of tampering methods with

*Corresponding author
Email address: p.a.johnston3@rgu.ac.uk (Pamela Johnston)

discussions of future research directions aimed at improving detection methods.

Keywords: video tampering, video synthesis, deep learning, video forgery

1. Introduction

The synthesis of convincing fake video content has increased recently due to the development of intelligent models [1, 2, 3]. Selective modification of image content has been possible for some years, but the application of similar techniques to video has been too labour intensive to see mass use. If each frame in a video is treated as an independent image, there are simply too many images to process efficiently. This has changed with increased computing power and the advent of deep neural networks. Deep learning techniques have seen great success in many applications recently. Generative Adversarial Networks (GANs) in particular have been used to alter source video: to re-enact human facial expressions [4], change the weather [5] and to apply face-swapping [6]. Human facial re-enactment is a relatively new but common area of research where a simple, talking head is visually altered to mimic the facial expressions of a second actor [4, 7, 8] or to match a different audio track [9, 10]. This may have innocent applications such as re-dubbing a film in a different language or creating new movie scenes using old footage of an iconic actor, but it can also be used to produce convincing fake content. In some circumstances, fake content is convincing enough to reliably fool human eyes. The authors of [8] even found that human viewers performed little better than random guessing when trying to ascertain whether facial re-enactment footage was authentic or synthesised. A deep neural network, however, could distinguish between the authentic and forged footage with ease.

Research into data-driven machine learning has also prompted the gathering of large image and video datasets such as ImageNet [11], Youtube-8m [12] and CelebA [13]. These datasets are a valuable resource for further research into convincing image and video forgery and in some cases, [8], a library of resources available for use in tampered datasets. The influence of these datasets has

led to an increase in the application of deep neural networks to tampering. Spatially localised changes in video footage, such as face swapping, can change the entire context of a news story or film and can have repercussions for the people portrayed. Already, videos which have been cleverly edited to change the context of what was said by influential people have gone viral [1] . If that can be done with unsophisticated editing techniques, it is worth considering what more could be achieved with modern techniques.

There are already a number of recent surveys which review tampering detection methods [14, 15, 16, 17, 18]. Tampering detection methods are broadly categorised as active or passive, with more focus on passive tampering detection methods. A review of passive tampering detection in video is provided in [16, 18] and, more recently, [14]. Inter-frame tampering detection is covered in [15]. Pandey et al [17] cover tampering detection through noise in images as well as video. There are, however, far fewer reviews on tampering itself. This paper aims to help redress the balance. The work in [19] provides an overview of personation, specifically how a person's likeness in appearance and voice can be forged in videos either physically or digitally. However it is important to objectively catalogue current known techniques that can be used for video tampering in order that they can be identified and, ultimately, detected or countered. Many detection techniques are explicitly tailored to specific tampering methods. For example, the authors of [8] trained a deep neural network to detect their own video content changes in order to assess the quality of their content-altering techniques; [20, 21] focused on inter-frame tampering; [22] created tampered sequences using established in-painting techniques [23, 24] to assess their detector. All of these techniques worked well, but all of them required prior knowledge of the *type* of tampering. As tampering methods multiply, it becomes important to fully assess new detection methods, and to appreciate which types of tampering

[1] "Video of Barack Obama speech circulating on the Internet was edited to change his meaning": https://www.politifact.com/truth-o-meter/statements/2014/jun/23/chain-email/video-barack-obama-speech-circulating-internet-was/ Accessed 2019-1-24

techniques they can feasibly detect and which they are blind to. This review exposes new research directions by cataloguing known tampering algorithms to aid development of automated, universal detection techniques.

Wang and Farid [25] noted that, at the time of their 2007 publication, there were very few video tampering detection techniques. This is no longer the case, however, many published techniques, specific to particular types of tampering or source authentication, were assessed on proprietary datasets which remain unreleased [21, 26, 27, 28, 29]. In some cases, [30], datasets are detailed sufficiently in the literature so that they can be replicated, provided the sequences used for dataset synthesis are available. This serves to evidence the fragmentation of the tampering detection field. In reality, a tampered video may be subject to a variety techniques, including combinations. For tampering detection to be effective, individual detectors must be analysed and matched with an appropriate type of tampering. In order for that to happen, we must review and differentiate types of tampering.

Here, we catalogue and analyse the current trends in digital video manipulation techniques from simple edits to fully synthetic video. This allows for work towards a method of universal video tampering detection. We list available tampered datasets and identify their potential challenges. We thoroughly examine problems in dataset gathering and dissemination, including challenges created by compression.

This review opens up a new field of research in the form of tampering classification. If video tampering is, by its nature, designed to be invisible to human eyes, then tampering classification will necessarily be done by algorithms and machines. As a first step in this process, we analyse the current classes of video tampering. Future tampering detection methods may maximise their impact by identifying and targeting tampering classes instead of individual algorithms. The purpose of this paper is to provide an in-depth analysis and review of existing methods of digital video tampering, to understand the current state-of-the-art, how it may progress in future and how this can be used to inform future development of tampering detection techniques. Our contributions are

as follows:

- A thorough examination of how video tampering techniques have been categorised in the past and how these have influenced development of tampering detection.
- An in depth, original review of how video tampering has evolved in recent times with discussion of the latest deep learning techniques.
- A qualitative evaluation of existing tampered video datasets. This includes critical analysis of large, tampered image datasets, which is used to assess the challenges associated with creating and distributing video datasets and propose effective methods for overcoming these.

The paper is structured as follows: Section 2 critically analyses traditional types of video tampering and how these have been augmented by the latest developments to form a spectrum of video tampering from fully authentic video through to fully synthesised video; Section 3 examines some state-of-the-art tampering/anti-forensic detection methods and discusses how these reveal underlying information about datasets used for training; Section 4 examines existing tampered video datasets and highlights good lessons that can inform future dataset compilation; Section 5 concludes the paper and gives new research directions.

2. Overview

In [14] video tampering was defined as "a process of malicious alteration of video content, so as to conceal an object, an event or change the meaning conveyed by the imagery in the video". Similarly, [31] described image forgery as "the digital manipulation of pictures with the aim of distorting some information in these images". In this paper, video tampering is regarded as any technique which is intended to produce manipulated, photo-realistic content using authentic sources. There is no defined limit as to when authentic video becomes tampered video. Similarly, malicious intent is difficult to quantify,

and so tampering detection and video authentication techniques must focus on forensic analysis, providing objective localisation of inconsistencies within digital footage that may imply content alteration.

It is important to note that this paper examines only *digital* video tampering: video content can also be "staged" whereby the video file is an authentic record of events, but the events themselves were contrived or unnatural during filming. Detection of staged video involving natural ballistic trajectories is examined in [32], and [19] details how plausible audiovisual personation is achieved in front of the camera. Digital video forgery can take a number of forms [14] and Figure 1 gives an overview of the classical interpretation.

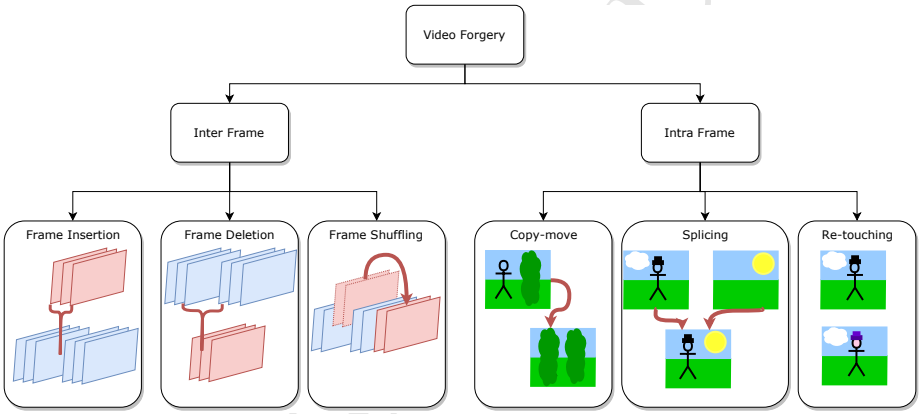


Figure 1: Traditional video forgery categories

In the past, video tampering methods have been simply classified as inter- or intra- frame [14, 15] (Figure 1). The terms inter- and intra- frame primarily distinguish temporal tampering from spatial tampering. Inter-frame tampering is performed on a sequence-level: the pixels of individual frames are unaltered, but the sequence as a whole is changed. Intra-frame tampering is performed on a pixel-level: some spatial regions are altered, but alterations temporally correlated to form a convincing forged region. The term "inter-video tampering" can also be used to describe the merging of content from two different videos [33]. Traditionally, this has been a form of splicing, where chroma-keyed objects taken from one sequence are inserted into another, as in [34]. Recent develop-

ments, however, mean that convincing synthetic regions [4, 6, 35] or even whole videos [2] can be synthesised automatically from authentic content. This development means that we must now consider different levels and categories of video tampering. It is important to be aware of the different categories because video tampering is designed to be invisible to human eyes, and detection techniques often address only one type of tampering.

2.1. The spectrum of video content

The current field of video tampering may be viewed as a spectrum, as in Figure 2, where different types of video tampering are ordered according to potential to deviate from authentic source. Whereas the traditional view in Figure 1 provides only two categories of video tampering, the spectrum in Figure 2 demonstrates that there are now multiple ways to produce convincing, falsified content. This distinction is important because detection methods often address one particular type of video tampering such as object forgery or inter-frame tampering. In [14] tampering detection methods are categorised as recompression, inter-frame forgery or region tampering detection. With current tampering techniques, the distinction is less clear cut. Moreover, multiple tampering techniques can be applied to the same sequence.

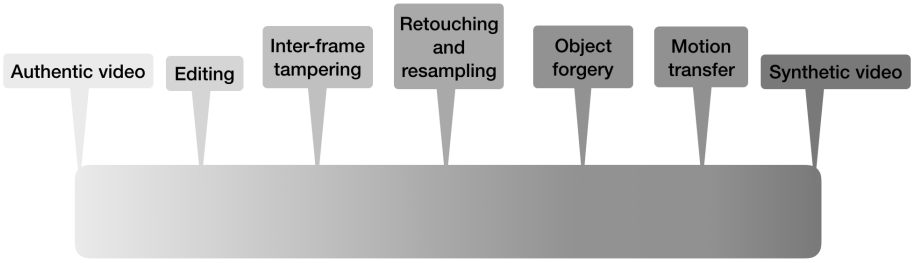


Figure 2: Video tampering spectrum

Figure 2 summarises the current categories of video tampering. Video editing compiles single camera shots into full films complete with scene cuts. Although clever editing may change the context of a video, scene cuts are not

usually deliberately concealed. Video clips of maliciously edited content exist in mainstream media and are surprisingly effective at disseminating misinformation through social media. Traditional inter-frame tampering, where edits are concealed, may reorder events or even remove or insert events into the timeline, but its content-altering effects are self-limiting. Retouching temporally or spatially upscaled content, or applying global filters to improve perceptual quality may affect every pixel in a video sequence and can cosmetically alter content. Retouching can also be applied to specific regions. Intra-frame tampering and other object forgeries such as inpainting can alter content and context, as can motion transfer. Finally, fully synthetic video or synthetic regions can be produced. Unlike historical animations, the synthetic content of today looks convincingly realistic. The following subsections 2.2 to 2.6 detail examples from each of these types of video manipulation.

Table 1 shows how motion transfer and video synthesis techniques have become common in recent years and demonstrates how methods of evaluation remain relatively underdeveloped. Evaluation techniques are difficult to define since there is no pre-defined ground truth for tampered video data. Every new method can be assessed qualitatively. Methods which seek to imitate authentic video, such as frame interpolation, can use full reference quality metrics such as SSIM and PSNR. As can be seen in Table 1, video manipulation methods use user studies to evaluate their output or simply publish examples of their methods for future evaluation. However, even user studies can vary. Some ask users to classify frames as tampered or authentic. Some request a user preference between the published method and other, similar methods. In a related field, image inpainting evaluation techniques are reviewed in [36] and these can all be used to assess the spatial features of inpainted video or indeed, any form of tampering which affects individual frames. No-reference video quality assessment is a large and open field and although we do not cover this here, we point to this field to at least partially inform on tampered video evaluation.

Table 1: Video Tampering and Evaluation Methods: Qual= qualitative analysis; PSNR=Peak Signal to Noise Ratio; SSIM=Structural SIMilarity; UP=User preference to previous methods; UR=User comparison with real video; Rel=Released Sequences

Reference	Year	Type of Tampering	Qual	PSNR/ SSIM	UP	UR	Rel	Other
ETS [24]	2004	inpainting	✓					
Ha et al [37]	2004	frame interpolation	✓	PSNR				
Patwardhan et al [23]	2007	inpainting via temporal copy-move	✓					
Wexler et al [38]	2007	inpainting, frame interpolation	✓				✓	
Shih et al [39]	2011	object forgery	✓				✓	
SULFA forged [40]	2012	object forgery	✓				✓	detection
SULFA supplemental [41]	2013	object forgery	✓				✓	detection
Newson et al [42]	2014	inpainting	✓				✓	
Ardizzone and Mazzola [33]	2015	copy-move	✓				✓	
Ebdelli et al [43]	2015	inpainting	✓	PSNR			✓	
Lotter et al [44]	2015	frame prediction	✓					error
Dar and Bruckstein [45]	2015	frame interpolation	✓	PSNR				
Face2Face [7]	2016	motion transfer	✓				✓	
Le et al [1]	2017	inpainting	✓				✓	
Suwajanakorn et al [4]	2017	motion transfer	✓					
Liu et al [5]	2017	style transfer	✓				✓	
Niklaus et al [46]	2017	frame interpolation	✓	PSNR				
MoCoGAN [47]	2017	motion transfer	✓			✓		ACD
Walker et al [48]	2017	frame prediction	✓					Inception
FaceForensics [8]	2018	motion transfer	✓			✓	✓	detection
Recycle-GAN [3]	2018	video synthesis	✓		✓			✓
Wang et al [2]	2018	video synthesis (sketch)	✓		✓			
Chan et al [35]	2018	motion transfer	✓	SSIM				LPIPS
Jiang et al [49]	2018	video synthesis (blurred image)	✓	PSNR				
Wang et al [50]	2018	video synthesis (smile)	✓		✓			
Xiong et al [51]	2018	video synthesis (time-lapse)	✓		✓	✓		
Babaeizadeh et al [52]	2018	frame prediction	✓	✓				
Zhao et al [53]	2018	frame prediction	✓	PSNR	✓			ACD
SCGAN [54]	2018	video synthesis (human pose)	✓		✓			pose eval.
SDC-Net [55]	2018	frame prediction	✓	✓				
Cai et al [56]	2018	frame prediction/interpolation	✓	✓				Inception

2.2. Editing and Inter-frame Tampering

Editing and inter-frame tampering both change the order of the frames in the video without changing the contents of each frame. In the case of editing, the goal is to turn a series of single camera shots into a coherent story. Clever edits can be used to turn innocent footage into propaganda,[2] but scene cuts are not hidden and such videos are not above suspicion. In inter-frame tampering, the goal is to *invisibly* remove, re-order or alter events.

Edits in inter-frame tampering are deliberately concealed so as to be invisible to the human eye. Detection of visible scene cuts in video has been studied extensively so that key frames can be identified for efficient compression and or used to condense/index the sequence [57]. Invisible scene cuts are studied in the context of inter-frame tampering detection [58, 59, 60].

Such is the theoretical simplicity of generating an inter-frame tampered sequence, that many tampering detection methods, such as [20, 21, 61, 62] generate their own datasets from single-camera video sequences such as SULFA [40] or Derf's media collection [63] or even film their own sequences as in [29]. SULFA [40] replicates single camera sequences as obtained from CCTV footage, and, therefore may be representative of the most likely application of inter-frame tampering: altering CCTV evidence. Derf's media collection [63], on the other hand, provides publically available uncompressed sequences and allows researchers complete control over the forensic history of synthesised tampered sequences [62].

It remains unclear how widespread inter-frame tampering is in the wild because, if it is done correctly, it will be undetectable by human eyes and remain above suspicion. Meanwhile, it is important that synthesised datasets are as high quality as possible. In creation of inter-frame tampered datasets, [29, 30, 64]

[2] "Israeli army edits video of Palestinian medic its troops shot dead to misleadingly show she was 'human shield' for Hamas", The Independent, https://www.independent.co.uk/news/world/middle-east/gaza-protests-latest-idf-condemned-edited-video-angel-of-mercy-medic-razan-al-najjar-a8389611.html

simply removed pre-determined frame numbers from each sequence, and it is unclear if this caused visible effects. In [62] frame addition and removal was limited to the beginning of each sequence, but again it is unclear if the additions were visible: simply reversing the sequence from the point of tampering may effectively locally conceal the edit. Recent developments in video quality assessment mean that temporal glitches in video can be objectively quantified [65] and also smoothed [66] to achieve temporal consistency. Future datasets for inter-frame tampering can use this to improve such that inter-frame tampering techniques can be deployed in the wild.

As noted in the review in [14], many inter-frame tampering detection methods suffer from limitations which are often related to consistencies within the dataset which may not translate to other video data. These consistencies are often related to video compression. The authors of [67] note that some tampering detection techniques are tied into the fixed Group of Pictures (GOP) size, commonly used in MPEG2 [68] to minimise error accumulation due to non-integer frequency domain transforms. Later video compression standards, such as H.264/AVC [69], use integer-based transforms so error accumulation drift between encoder and decoder is no longer an issue, and therefore key frames are used only as access points into the stream or efficient compression of cut scenes. Moreover, sequences compressed using [69] no longer exhibit visible evidence of key frames.

Although inter-frame tampering detection is widely studied in the literature, effects similar to inter-frame tampering can be achieved using a spatio-temporal copy-move. Rather than replacing complete frames in the sequence, only partial frames containing motion or objects to be concealed are replaced. With a static camera and consistent lighting, this is visually effective, and video edits prove near invisible to the naked eye. Indeed, some sequences which initially look like inter-frame tampering [41] are actually spatio-temporal copy-moves, as can be revealed by examining pixel-by-pixel difference between the authentic and tampered sequences (see Figure 3d).

2.3. Retouching and Resampling

Retouching involves adjusting pixels within an image using transforms or filters applied to the pixels themselves which may only have a low-level interpretation of video content. As the name suggests, retouching is less invasive to content than other types of forgery but may still change the context of a video. Moreover, retouching can be used on tampered video specifically as an anti-forensic device.

A retouching function R can affect specific pixels according to a binary mask, M:

$$V_{retouched} = R(M \odot V_{original}) + ((I - M) \odot V_{original}) \tag{1}$$

Here, I represents a matrix of ones and all matrices have equal size. Retouching can also be applied globally as in:

$$V_{retouched} = R(V_{original}) \tag{2}$$

Colour correction methods such as those available in Adobe After Effects may normalise lighting in a sequence of shots taken on different days under different weather conditions to create a convincing narrative. Similarly, colour grading can also be used to add effects or make video filmed during daylight hours appear to have been filmed during twilight. A typical colour correction model works by adjusting the histogram of colour over a specified region, however the authors of [70] found that gamma correction (a form of colour correction) was particularly difficult to detect using a deep neural network. Where median filtering and Gaussian blurring could be detected reliably with over 91% accuracy, detection of gamma correction was only 57.6% .

Compression is often a necessary part of video processing but it can also be used as an anti-forensic method. Video compression standards such as [68, 69] reduce video file size with no explicit understanding of video content. Compression has been found to reduce the efficacy of tampering detectors [8, 41, 64, 71] and has also been shown to reduce the classification accuracy of convolutional

neural network (CNN) based classifiers [72]. The authors of [8] found that video compression [69] reduced the accuracy of deep neural networks trained to detect human facial re-enactment. Of seven different forgery detectors tested, the Xception network [73] was the most robust against compression, achieving 87.81% accuracy, compared with 99.93% accuracy on uncompressed sequences. Other methods [74] performed less well for forgery detectors on the compressed dataset with performance for some [75] dropping as low as 55.77%. This may be attributed to the depth of the Xception network. The authors of [71] also account for compression in their SYSU-OBJFORG dataset. In benchmarking it using seven common steganographic features, they, too, found a drop in accuracy on the reduced bitrate video. While their ensemble-based detector achieved precisions in the range 78.9-93.15%, this reduced to 61.85-79.34% when bitrate of the video data was halved. Halving height and width of the video sequences also reduced precision to the range 73.02-90.28%, which was not as significant as bitrate reduction. In detection of frame deletion, it was found in [64] that an SVM conditioned on uncompressed data to detect dropped frames did not perform well on compressed data taken from YouTube, with accuracy dropping below 37%. It is clear from this that standard video compression can reduce some of the features associated with tampering. This is most likely down to the way video compression quantises data in the frequency domain.

Compression artifact removal is another example of retouching and methods such as [76, 77, 78] have been applied to JPEG images. The authors of [79] used a deep residual network to reduce artifacts in a BPG [80] compressed frame. More recently compression artifacts have been removed in the video domain [81], where videos compressed using HEVC and H.264/AVC were retouched to increase Peak Signal to Noise Ratio (PSNR). PSNR is defined as:

$$PSNR = 20log_{10}(\frac{255}{\sqrt{MSE}}) \tag{3}$$

where MSE is Mean Squared Error, and it is a very common full reference quality metric where processed pixels are compared directly to unprocessed

pixels. It is useful for measuring pixel fidelity and is often used alongside the full reference quality metric Structural Similarity (SSIM) [82], but may not directly reflect perceptual quality. Although [81] achieved overall improvement in PSNR, it was unclear if this resulted in a gain in perceptual quality at specific bit rates. Given that some tampering detection methods, such as [83], actively utilise compression structures, methods which alter the underlying patterns of compression in video frames could be used as anti-forensics in the future.

Artificially upscaling video frame size [84], frame rate [85, 86, 87] or bitrate can be a form of video tampering. High quality video content is more desirable to consumers, and larger file sizes for the same film/footage are often indicative of higher quality, with bitrate often taking the place of quality in common parlance. Compression encoders utilise a specified bitrate, even if this means compressing existing compression artifacts. Bitrate upscaling can be done innocently as researchers seek to provide high quality, "uncompressed" datasets and either overlook or deliberately replicate compression artifacts in the pixels of mined data.

Artificially increasing the spatial dimensions of video has been commonly done historically as Standard Definition (SD) content is displayed on High Definition (HD) screens. More recently, super-resolution has evolved from an image enhancement technique to use within videos [84, 88, 89], and the metrics commonly used for evaluation are, again, PSNR and SSIM: full reference quality metrics. It is important for spatially upscaled video to demonstrate temporal coherence. The authors of [84, 88], also assessed the temporal coherence of their super-resolution sequences using a technique called "temporal profile". This is where single rows of pixels are viewed along with their temporal neighbours from different frames, and temporal inconsistencies or video "flicker" shows up as hard edges in the resultant image. The work in [65] is also a method of assessing temporal consistency.

Video deblurring [90] is another example of retouching and a dataset exists to facilitate the development of this [91]. The dataset was filmed using a GoPro camera at 240fps and then downsampled and blurred so that a non-blurred

ground truth can be supplied for each blurred frame, thus enabling deblurrers to be assessed using full reference quality measures such as PSNR. Super slow motion has recently become a strong field of research with many new techniques for temporally upsampling video [46, 85, 92] to create a slow motion effect in the absence of a high speed camera. Previously, upsampled video would simply involve frame repetition or averaging. The field of motion compensated interpolation improved upon this [37, 45, 46] so that interpolated frames were less obvious. The work in [37] is an early example of motion compensated interpolation, and the authors used block-based motion estimation similar to that used in video compression [68] to inform interpolation and create sophisticated intermediate frames. The frame rate upconversion algorithm was objectively assessed by downscaling some publicly available uncompressed sequences and then comparing the original sequence with the computed upscaled version using PSNR. In [45], the authors showed how their work in frame rate upscaling could be used to improve low bitrate video compression. In [46], a CNN was used to interpolate between frames. The authors obtained their training data from high quality YouTube channels and downsampled from 1080p to 720p in order to reduce the effects of compression. A user study confirmed that their interpolated frames were better than previous state-of-the-art. Objective assessment of results used sampled alternate frames from a popular YouTube video and used PSNR to compare interpolated frames with actual frames. In [85], multiple frames were synthesised between two authentic frames using a CNN trained on high frame-rate (720p, 240fps) video from YouTube and a high frame-rate dataset [93]. The synthesised frames were assessed using high frame-rate video and it was found that PSNR between interpolated frames and ground truth was improved upon previous state-of-the-art. As noted in [94], inpainting or video completion (Section 2.4) can be used to resample a video, and entire frames inpainted.

Retouching might be one step of many in tampering, and although it does not necessarily alter context, it can be used to make tampering detection much more difficult. Countermeasures for anti-forensics are well studied in the literature

(a) Auth. 34 (b) Auth. 236 (c) Tamp. 236 (d) Difference

Figure 3: An intra-frame tampering example from [41]. 3a and 3b show authentic content. 3c shows the spatio-temporal copy-move and 3d shows the difference between 3b and 3c

[70, 74, 95], and datasets can be generated easily. In [74], a CNN was used to classify an image in terms of its anti-forensic processing. The labels used were: original (no processing), Gaussian blurring, additive white noise, median filtering and resampling. The CNN accurately detected the presence of each process over with over 98% accuracy using only the green colour channel. A new type of convolutional layer was designed to prevent the network from learning typical image features. The authors of [95] showed how their median filter detector could be used to localise median filtering within a spliced image and hence localise image tampering. Although retouching does not always correlate with tampering, localised retouching can be a strong indicator of splicing or other object forgery.

2.4. Intra-frame Tampering

Intra-frame tampering is where spatial content of individual frames is changed, that is, individual objects are added or concealed/removed. Intra-frame tampering is also known as "region tampering" [18] and applies equally to video and still images, although the video application is more complex. Care must be taken to ensure that spatial tampering across individual frames is coherent and does not cause visual jarring in the video. Intra-frame tampering methods in images were classified as spatio-temporal copy-move, splicing and retouching in [96], but here we discuss retouching separately (Section 2.3).

A spatio-temporal copy move can be defined by:

$$V_t^{Lj} = ((I - M) \odot V_o^{Lj}) + (M \odot V_o^{Lk}) \tag{4}$$

where I is the matrix of ones, M is a binary mask to localise tampering, V_o^{Lj} is an authentic sequence of L frames starting on frame j, V_o^{Lk} is the same sequence but starting on frame k where $j \neq k$. The frames/mask can be re-aligned or cropped so that any object or region of pixels from any spatial or temporal location can be copied to any location. V_t^{Lj} is a tampered video sequence:

$$V_t^{Lj} = [v^j, ..., v^{j+L-1}] \tag{5}$$

In spatio-temporal copy-move attacks, all the data used in the video forgery V_t comes from within the same video sequence V_o. For example, complete objects from frame k in the sequence are inserted into frame j using mask M. Figure 3 shows an example. This is similar to image-based copy-move where the pixels involved in the tampered region come from within the image itself. Although this reduces the range of potential content, it helps to minimise differences between legitimate and tampered regions. There is less need to alter the colour histogram or adjust the frame rate to make tampered content consistent with authentic content if both share the same source. Using a copy-move attack, objects can be added to a sequence by adding foreground objects, or concealed/removed from a sequence by duplicating background regions from within the same frame or from within a different frame in the same sequence.

Some versions of copy-move attacks simply duplicate a still background region, [40], and these can be detected with relative ease by high coherence or abnormally low motion within the tampered region, [41]. Other methods [41] duplicate an entire spatio-temporal region, and this is more difficult to detect. Although duplicates can be detected by matching copied region to original data, this becomes more difficult in the presence of compression [41]. A copy-move attack can be detected in images by identifying and locating duplicated regions, and this has been done using search based on brute-force pixel matching, region

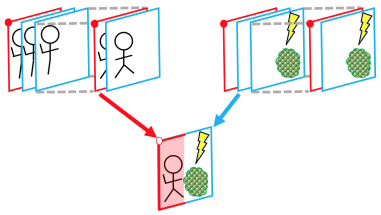


Figure 4: Forging a video (best viewed in colour), Red borders/dots indicate key frames in the sequence. Blue borders (no dots) indicate predicted frames. The hybrid frame is shaded red where the pixels have come from a key frame, and unshaded where the donor frame was a predicted frame

matching or key point matching [96, 97]. While this type of copy-move detection is feasible in images, video adds another dimension and searches become an order of magnitude more complex. Previous video inpainting attempts such as Temporal Copy-Paste (TCP), where identical pixels are used frame after frame to conceal an object within a video [40], or Exemplar-based Texture Sythesis (ETS) [24] were detected by [22] where detection was based on correlation between adjacent frames which was either too strong or not strong enough to be authentic video.

Splicing is an extension of spatio-temporal copy-move. In a splicing attack, two sets of pixels from different sources are combined as in Figure 4. The sources can be videos or even still images (as shown in Figure 5). Equation 6 defines splicing:

$$V_t^{Lj} = ((I - M) \odot V_{s1}^{Lj}) + P(M \odot V_{s2}^{Lk}) \tag{6}$$

Where video sequences are defined as in Equations 4 and 5, $s1$ means sequence 1, $s2$ means sequence 2 and P is an optional processing step which can be applied to aid blending between different source videos. The frames/masks of the two sources can be re-aligned or cropped so that any object or region of pixels from any location from source 2 can be pasted into any location in source 1.

Splicing has large potential for context-changing edits because two entirely

(a) authentic (b) tampered

(c) authentic (d) tampered (e) mask

Figure 5: An example of spliced content. 5a and 5b are from VTD [98] and the spliced content (a picture on the wall) comes from a static image. 5c, 5d and 5e come from D'Avino et al [34] where the spliced content comes from a chroma-keyed video

different subjects can be spliced together. Any source sequences involved can be retouched (see Section 2.3) using colour correction or temporal synchronisation before or after a video splice in order to visually camouflage spliced content, or even to launder the splicing operation to make it undetectable to existing forensic tools.

Copy-move and splicing are also known as "object forgery" [71] because they often involve removing or adding complete objects to videos. Introduction of an object to a video can be done using chroma-keying techniques, as in the field of video special effects [99]. Chroma-keying requires filming against a single colour background under specific lighting conditions to facilitate segmentation of foreground objects. An example of object forgery using chroma-keyed sources is shown in Figure 5d. Other segmentation methods such as [100, 101] may be used in place of chroma-keying so that foreground objects can come from any sequence without the need for special green-screen filming. The authors of [100] applied segmentation to the optical flow of videos in order to distinguish

foreground and background objects in sequences with moving cameras. In [101], segmentation was achieved using supervoxels, using spatiotemporal uniformity in pixels to group them into voxels and supervoxels to represent different objects in the sequence. Masks produced by [100, 101] could be used in place of specially filmed green-screen sequences, thus rendering any video susceptible to use in object forgery.

Inpainting or video completion [23, 38, 42, 43, 94] allows removal of objects from an image by interpolating remaining pixels to conceal a "hole" left by a removed object or corrupt section of video. It is useful for error concealment when streaming video over an unreliable channel and can be used to restore old film, but it can also be used to deliberately remove objects or even frames from a sequence. Video in-painting techniques were surveyed recently in [94], where it was noted that many methods of video in-painting rely on patch completion where the missing spatio-temporal volume is filled using small patches from within the same video sequence. This is evident in early video in-painting methods such as [23, 38]. In [23], static background and dynamic foreground were assumed, thus highlighting one of the challenges associated with video completion: motion. This was handled by first registering or aligning the frames of the sequence. Background mosaics of the video sequence were then constructed by removing all non-static objects, and foreground mosaics contained all the moving objects. Missing data was then inpainted by finding close matches in the mosaics and interpolation with texture synthesis between the matched segments. The authors of [38] used space-time volumes of 5x5x5 pixels taken from other areas of video sequences to fill in the space left behind be a removed object. Motion was accounted for by representing each pixel not only in terms of its RGB components but also two components based on the derivation along the x-, y- and t- dimensions. This method also allowed temporal and spatial upsizing as spatio-temporal holes had varying dimensions in the spatial and temporal axes. More recently, the work in [42] realigned source patches to create closer matches and less warping of synthesised video content. Initial values for missing pixel data were also explicitly defined in [42].

The assessment of inpainting quality in the image domain was critically reviewed in [36], and user survey to assess the visibility of inpainted regions remains the gold standard. The authors also noted that Video Inpainting Quality Assessment remains an important, open area of research. Indeed, in the absence of an accepted video completion quality assessment, authors [1, 42, 43] simply publish videos of their inpainting techniques applied to standard sequences online and [94] notes this as a trend. The authors of [43] also provided their original sequences along with defined masks so that future inpainting techniques can be applied to precisely the same data for comparison. The existence of these inpainted sequences provides a good source of data for video tampering detection research.

Inpainting can be used in conjuction with spatio-temporal copy move to create complex forgeries. An early example of this can be found in [39] where the authors changed the winner of a 100m race. The authors considered the video as a series of layers. They applied in-painting using unoccluded areas of background and interpolated/sampled the motion of forged runners to make them move slower/faster relative to other objects in the video. While individual frames taken from the forged sequences looked visually convincing, full video sequences are not currently available for full analysis. Indeed, assessment of tampered video remains an open problem, one which the authors of [39] suggest is best tackled by forgery detection methods. Although the subject matter of this video was somewhat ambitious for its time, and the authors explicitly target the field of video special effects, it gives a good idea of how tampering can be used to court controversy.

2.5. Style and Motion Transfer

Style transfer is a new method of image and video manipulation which has been facilitated by the advent of Generative Adversarial Nets (GANs), which were first established in [102] and extended to conditional GANs in [103]. Style transfer can completely change the context of an image or the subject of a video. It is strongly related to motion transfer because the resultant video is

a combination of motion from one source video and content or subjects from another. Combining the two can be viewed as a style transfer when the style of the content source is mapped to the motion source or it can be considered motion transfer when motion is mapped to the content source.

Examples of style transfer in the image domain include [104] where features from one object are mapped to a similar object: a scene can be changed from a summer scene to a winter scene; a horse can be exchanged for a zebra [105]; Google Street View House Numbers can be translated into MNIST-style digits [5] and evaluated using accuracy on a CNN trained to classify MNIST. Examples in the video domain include motion transfer [4, 7] as well as style transfer.

An example of a conditional GAN used to perform style transfer can be found in the seminal Pix2Pix [104], which performs image to image translation. A GAN consists of a generator network and a discriminator network. In Pix2Pix, the generator network maps an observed image and a random noise vector to a generated image. The discriminator network then uses both the mapped image and the observed image to classify the mapped image as an example from the authentic dataset or one from the generator. Authentic examples given to the discriminator dictate the "style". This architecture is distinct from a non-conditional GAN where the discriminator network sees only the mapped image. The authors of Pix2Pix noted that they could achieve very good results based on small datasets of only 400 authentic images and so the GAN can be trained for a multitude of applications. For example, the input image can be a sketch or a semantic segmentation mask and the mapped image can be photorealistic, or vice versa; daytime scenes can be mapped to night; the mapping process can even perform inpainting or background removal. The versatility of [104] has also spawned further applications in the video domain including [3, 47].

Motion transfer is similar to style transfer where the motion of one object is passed on to another object. Early applications were mostly specific to human facial re-enactment such as lip synchronisations and expression translation between talking heads [4, 7]. Thies et al [7] presented the first real-time facial re-enactment system that used only RGB as input. The method used authentic

frames from a target video and transformed them to match the facial expressions and mouth motions from a source video. In [4], the authors added video re-timing for realistic head motion to fit the context of the spoken word and used a recurrent neural network (RNN) trained on many hours of footage of the particular subject to transform an audio track into mouth shapes. While [4] was not real time, and required many hours of video footage to train the RNN, it was capable of producing a representative video from audio and stock footage, whereas [7] required video for both source and target. More recently, motion transfer has been achieved using models based on style transfer.

MoCoGAN [47], used GANs in a similar way to Pix2Pix [104]. Content and motion were treated independently in MoCoGAN and video sequences expressed as:

$$Z_i = Z_c \times Z_m \tag{7}$$

Every frame in Z_i has a content vector, Z_c, and a motion vector, Z_m, associated with it. In order to perform motion transfer, the content of one sequence was substituted with the content of another. The architecture consisted of two distinct discriminator networks: one to classify real and generated images (or frames), and one to distinguish real and generated video. The video discriminator was responsible for smooth video generation. Similarly, there were two connected generator networks: one to generate motion, the output of which was used to condition the content generator which produced video frames. The motion generator network was a recurrent neural network (RNN) which modelled motion through time. Motion content could also be extracted from a different sequence and hence motion can be transferred between two similar videos. The authors of [3] also applied motion transfer to videos, successfully replicating lip motions. Because [3, 47] are both based on style transfer, they can also be used to create photo-realistic synthetic video from semantic segmentation masks (Section 2.6).

The assessment of GAN-produced images and videos remains an open prob-

lem. In [5, 104], translated images were objectively assessed using the accuracy of CNNs pre-trained on authentic images in the output-style classifying the translated images. It was found that the CNNs classified the translated image of [5] with more than 90% accuracy. Image translation methods from [5] were also applied to some street driving video sequences, and qualitative analysis of the results showed a convincing, low frame rate video where the weather had been translated from sunny to snowy or the lighting mapped from day to night. The authors of [106] applied neural style transfer to photographed objects spliced into images of paintings, thus reducing the visibility of the tampered object. A user study found that their edited image set achieved similar user scores to an unedited image set meaning people could not reliably localise such processed image edits. Although [5, 104, 105, 106] show a method to alter image content, they do not assess whether there is a counter method which can detect these alterations. Since all of these methods employ the use of GANs, it is implicit that there already exists a network which has been trained to discriminate between authentic examples of the style and synthesised content, but due to GAN convergence, this network may not be optimal for detection. In the video domain, the authors of [4, 7] have released examples of their work to the public. In [9], the authors used a user study to compare their audio-to-video speech synthesiser to both a previous model and a motion capture solution. The study showed that their work advanced the state-of-the-art as their examples were preferrable to human eyes when compared to previous speech synthesis, but not when compared to motion capture generated video. The authors of [8] performed a user study on their tampered dataset and found that, when asked to differentiate between tampered and authentic videos, humans achieved no better than random guessing. Generic motion re-enactment and video generation has also been studied in [47] however state-of-the-art is not yet of a standard where such tampered videos are high quality content.

Image to image style translation can be applied to video frames to universally change the overall context of the video. Complimentary work can be found in [66] where the authors examined the removal of flicker from a sequence of frames.

They specifically aimed to allow the use of image style transfer on individual frames to produce a temporally coherent video sequence, independent of the style transfer method. In [35] the authors used style translation on videos to synthesise video content of people performing dance moves they had never done. Pose estimation was used as an intermediate step. A conditional GAN was trained to map a stick-man pose estimation to a photo-realistic video frame using the previous frame to condition the GAN. They then applied a spatio-temporal smoothing to generate convincing videos that showed a target actor dancing in a manner defined by a source actor from a different video. The video sequences were assessed by extracting a pose from the mapped sequence and comparing it to the pose used to generate the sequence, and manual qualitative analysis of temporal qualities including some publicly released sequences. The authors conceded that there were still a number of challenges to overcome in this field, such as loose clothing and cluttered background, but it is easy to see that motion transfer can already be convincingly applied to human faces and bodies.

2.6. Photo-realistic Synthetic Video

Although purely synthetic video in the form of animation has been around for a long time, more recently synthetic video has been generated which is so photo-realistic that it could be mistaken for authentic, filmed content. In this section, we examine the most recent techniques in photo-realistic video synthesis and discuss their evaluation. Although video synthesis is not explicitly tampering an existing video, full convincing, photo-realistic video synthesis has the potential to be just as damaging as motion transfer or inpainting. It is important to examine it with a view to detecting it as a future research direction. Current trends in top international conferences on computer vision show that video frame prediction is a strong trend.

A short video sequence was extrapolated from the motion blur of a single image in [49]. The authors noted that the main challenge of this is temporal ordering. While the central frame of the synthesised sequence corresponds to

the de-blurred image, the motion of individual objects in frames before and after is ambiguous. The authors proposed a pair-wise ordering invariant loss to aid convergence of their CNN, which was based on pairs of frames at an equal temporal distance from the middle frame. Although the de-blurring aspect of the technique improved on the previous method [91] for moderate blur, evaluation of the short synthetic sequences proved difficult. The ambiguity in temporal ordering could be resolved when the process is constrained to temporal super-resolution. Video generation from a single image was also covered in [50] where Wang et al detailed a method to produce a short photo-realistic video of a smile from a single aligned face image. The method use a series of conditional Long Short Term Memories (LSTMs) to produce a sequence of facial landmarks moving from a neutral expression to a smile. A network similar to [104] was then used to translate the facial landmarks into a realistic video. A comparative user study found that the resulting sequences looked more realistic than a previous method, but the authors noted that it was difficult to evaluate such a method as there were no directly comparable existing methods. Both [49] and [50] extrapolated short synthetic video sequences based on a single image, and both noted challenges in evaluation. Xiong et al [51] produced short, realistic time-lapse videos of skyscapes, up to 32 frames from a single image using a two-stage GAN architecture. The first stage produced a sequence of frames and the second stage refined it to produce a coherent video. They also gathered a large dataset of real time-lapse videos from YouTube for the purposes of training. Again, there was no previous work available for direct comparison, but the authors were able to repurpose other network architectures to synthesise time-lapse videos. Evaluation was by user study where users were asked to identify the more realistic of two sequences. Although the proposed method outperformed all other synthetic videos, when comparing synthetic video with real video, only 16% of synthetic video tests were preferred.

In [44], a CNN, LSTM and deconvolutional neural net were used together to predict the next frame in a video sequence. The authors noted that natural images were much more challenging than simple moving circle animations, and

that, in predicting the next frame in a face rotation sequence, the network altered sufficient features so as to change the perceived identity of the face. The work in [52] followed this, using a variational autoencoder to predict the next 10 frames from a 10 frame sequence. The authors tested their sequences on [107] among others where they compared their predicted frames with ground truth frames using PSNR and SSIM. They conceded that assessing the quality of the predicted frames was difficult, and that the prediction yielding the worst PSNR was sometimes qualitatively the best. They also publicly released many examples of predicted sequences. In [108], the authors used a two stream structure and RNN to perform frame prediction. The authors of [56] viewed frame prediction as analogous to frame interpolation and fully synthetic video generation. They successfully produced short video sequences which interpolated between two frames as well as predicting short sequences given only the first frame of each sequence. Evaluation used PSNR and SSIM for interpolated sequences and Inception scores for generated sequences with no ground truth. Frame prediction was also covered in [53, 55, 109], and evaluation also involved full reference quality metrics. Although methods were evaluated using full reference quality metrics, there is no guarantee that frame *prediction* as opposed to interpolation will predict a frame that matches the original sequence, but it may yet produce a valid, realistic frame.

Some methods create synthetic video data specifically for machine learning datasets. In [110], a dataset of synthetic videos was created from motion capture data. The motion capture data was used to generate 3D models of human bodies which were combined with a texture map to add clothes and skin, and a static background image. The synthetic videos, rendered using Blender, were found to improve body part and foreground background segmentation. The Human3.6M dataset [107], includes some mixed reality videos which consist of a moving synthetic human model combined with a real video sequence. The real backgrounds included annotated occluding items so that synthetic human models could realistically interact with authentically filmed objects. Neither [110] nor [107] are specifically designed to fool human eyes, but instead intended

to aid development of human pose estimation and body segmentation. Rather than annotate thousands of frames of authentic video, the synthetic human model is already annotated. This is an example of a non-malicious application of synthetic video, although the detection of the synthetic parts is often trivial to human eyes.

Wang et al [2] have already synthesised coherent, photo-realistic video sequences of up to 30 seconds from semantic segmentation mask or pose model sequences. A GAN was used, and a discriminator part of the GAN used to classify the content as an authentic video or not, similar to MoCoGAN [47]. Using a discriminator in this way ensured temporally coherent video. The authors conceded that significant changes in an object's appearance is still a substantial challenge and that their model was also prone to colour drift over time, however a user study showed that [2] produced video that was preferable to human eyes than that produced by MoCoGAN [47]. SCGAN [54] also performed a user study which put their synthetic video content at a higher level of realism than MoCoGAN. Recycle-GAN [3] was also used to generate photo-realistic video from semantic segmentation mask sequences and assessed their method's accuracy by asking users to classify videos as synthetic or real as well as comparison with existing state-of-the-art. Users were fooled into thinking synthetic video was real 28.3% of the time. Quantitative results were also obtained using the Viper dataset [111] which supplies pixel-level segmentation masks for computer game scenes with a high level of realism.

Methods of evaluation for synthetic video remains an open field. It can be seen in Table 1 that the main methods include full reference quality metrics PSNR and SSIM as well as a variety of others such as Average Content Distance (ACD) [47, 53], Learned Perceptual Image Patch Similarity (LPIPS) [35] and tampering detection methods. Many of the techniques for synthetic video generation also utilise user surveys to assess the quality of synthetic video, and in many cases, evaluation is relative to previous related work (Table 1). It can be inferred from this that although current methods do not yet reliably generate video that is photo-realisitic enough to fool human eyes, improvements are

continuous and incremental. It is simply a matter of time before photo-realistic synthetic video becomes mainstream. As [8] showed, some techniques are already indistinguishable from authentic video for human viewers. This raises the problem that, in future, not only will humans be unable to detect tampered or even photo-realistic synthetic video, they will also be blind to whatever tampering technique has been applied. When this is the case, universal tampering detection systems will be required to fill the gap in human perception, and these must be developed urgently if detection systems are to keep pace with tampering methods. For this, datasets are required.

3. Image Tampering Detection

To advance the field of universal video tampering detection, it is vital to gather datasets of independent examples of video tampering techniques. In this section, we look at the lessons relating to tampered image datasets that can be learned from the application of deep neural networks to the problem of tampering detection.

As machine learning techniques come to the fore in tampering detection, the collation of large datasets to train and test networks becomes desirable. However it is important to realise that *any* consistencies within labelled classes may be exploited as features by deep learning techniques, including any features arising during dataset generation that are unrelated to actual tampering. In 2011 Torralba and Efros [112] discussed how bias is ubiquitous within computer vision datasets. Images from the same dataset exhibit characteristics specific to that dataset, so much so that a basic support vector machine (SVM) classifier trained to label a given image with its associated dataset achieved reasonable accuracy of 39% over 12 datasets. Each dataset has its own inherent distribution which may be irrelevant to the real world situation, and may be overlooked by human eyes. This problem is subtly highlighted by the advance of deep learning, particularly in the field of image forensics.

A good example of unintentional features comes in the CASIA2 TIDE dataset

[113]. This large dataset consists of 7491 authentic and 5123 tampered images which use splicing or copy-move techniques. The size of this dataset makes it an attractive option for deep learning and over 97% classification accuracy has been achieved by [114]. However, as noted in [115], compression applied to tampered images of the dataset differs from that applied to authentic images. Put simply, during dataset generation, tampered images were compressed twice, authentic images were compressed only once. There were also patterns in the colour space resolutions with tampered images more likely to have lower colour channel resolution. This means that classifying a CASIA2 image as tampered or authentic can be accurately achieved using features of compression, recompression and colour resolution. The recompression step may have arisen from the tools used to tamper the images, but it is independent of the tampering task itself. There is no reason that an authentic image cannot be innocently recompressed.

In [116], dataset weaknesses such as those in CASIA-2 were used as an explanation for the sharp drop off in CNN classification accuracy whenever the test images were compressed. Classification accuracy dropped from 97.44% on unprocessed CASIA-2 image patches to 68.11% when the images were compressed with JPEG quality factor 90, a fairly light compression. The authors proposed a means to circumvent this dataset flaw by extracting authentic *patches* from tampered images, however they did not report whether this reduced the drop-off in accuracy when the source dataset was compressed, nor did they report on a CNN *trained* using compressed image patches. Maliciously tampered images in the wild are not necessarily recompressed, and authentic images are not necessarily compressed only once. Tampered and authentic patches may be extracted from only the tampered data but only if reliable localisation masks exist to differentiate tampered and authentic pixels.

High levels of classification accuracy were also achieved by a deep neural network on the large rebroadcast dataset presented in [117]. This dataset comprises over 29000 images, half authentic and half rebroadcast in some way. Rebroadcast techniques included printing out and rescanning/photographing the images, screen grabs and screen photography. While some traditional techniques [118]

demonstrated poor accuracy on this new dataset, a CNN trained on 60% of the images and tested on the remainder achieved over 97% accuracy. In this case, recompression is very likely a necessary feature of retransmission, so features that emerged during CNN training are a true reflection of the real process of retransmission. One way to objectively assess this is to check the performance on a rebroadcast test set gathered independently. If a deep neural network exploits unintentional weaknesses inherent in a particular dataset, then the learning will not transfer well to other, similar datasets unless they exhibit the same features.

Table 2: CNNs for image anti-forensics detection

Reference	Detection of:	Dataset	Accuracy
Bayar and Stamm [74]	Gaussian blurring, additive white noise, median filter, resampling	proprietary	99%
Choi et al [70]	all combinations of Gaussian blurring, Median filtering	[119], [120]	>91%
Choi et al [70]	Gamma correction	[119], [120]	57.6%
Amerini et al [121]	double compression level	[122]	83.5% - 99.9%
Boroumand and Fridrich [123]	low-pass-, high-pass-, denoising- filters and tonal sharpening	[119]	>95%
Agarwal [117]	rebroadcast	public [117]	>97%

Table 2 shows how CNNs excel in detection of image anti-forensics. A detection or classification accuracy of over 95% is a common occurrance. Anti-forensics are methods designed to "launder" tampering and thus fool tampering detectors. Laundering techniques include general filtering methods such as compression, median filter and Gaussian blur. This field is emerging rapidly because large datasets can be synthesised with relative ease, and this makes it particularly appropriate for machine learning. Datasets such as BOSSBase [119], UCID [122] and Dresden image database [120] provide a large variety of unprocessed images to which known anti-forensic techniques can be universally applied and subsequently detected.

In [123], a CNN was trained to identify laundering techniques applied to an image. The laundering techniques, applied singly, were: low-pass-, high-pass- and denoising- filters and tonal sharpening. The authors first compressed the images of the dataset, then applied a single laundering technique and then

rescaled, cropped and recompressed the resulting images. Compression was JPEG with a Quality Factor (QF) ranging from 75-95. They achieved over 95% accuracy in identification of the laundering technique used regardless of the image compression level, provided the CNN was trained on images with a QF similar to that of the test dataset. The idea that a training dataset must be well matched to the test data in terms of compression is also supported in [72]. All of these high accuracies show that machines are adept at detecting patterns in visual data which are invisible to humans. This makes designing a dataset which is representative of the problem of video tampering but immune to unintentional side effects especially important.

4. Tampered Video Datasets

With so many different methods of tampering already available, and the field progressing at an unprecedented rate, it is important for tampering detection techniques to keep pace. Unfortunately this is challenging because there are few large, diverse tampered video datasets. Pandey et al [17] noted that tampered video datasets lag far behind tampered image datasets in terms of maturity. In this section, we examine existing video datasets and give recommendations for the design of new datasets. Table 3 provides a list of video tampering datasets, specifying the type of tampering applied and the size of the datasets.

A number of tampered video datasets already exist, but these vary both in terms of processing and parameters. In [34] the authors supply tampered video along with an explicit pixel level binary mask detailing the chroma-keyed addition. Some video tampering datasets come complete with original and tampered videos, thus providing a means to calcualte all masks and labels associated with tampering [41, 98]. This allows for tampering detection and localisation in spatial and temporal domains. It also allows for any differences in distribution between tampered and authentic sequences to be overcome by extracting authentic patches from tampered sequences. However, an accurate mask can only be extracted where videos can be synchronised and are identically processed

Table 3: Tampered Video Datasets

Name, Date, Ref.	Type of Tampering	Size	Details
VTD 2016 [98]	splicing; copy-move; frame-shuffling	26 tampered + related authentic	Distribution on YouTube means all videos affected by varying compression
FaceForensics 2018 [8]	motion transfer ([7])	1004 tampered x2	Taken from Youtube-8m [12], one set self-re-enactment, one set source-target translation
SYSU-OBJFORG 2016 [71]	object forgery	100 authentic, 100 tampered	No source for public download
D'Avino et al 2017 [34]	splicing	10 tampered	Binary masks provided, tampering is easily seen.
SULFA forged 2012 [40]	spatio-temporal copy-move	5 tampered	Static camera, background duplicated to conceal objects. Part of a larger database including untampered video for camera identification
SULFA supplemental 2013 [41]	spatio-temporal copy-move	10 tampered	Duplicate spatio-temporal regions to conceal/introduce objects
Lin et al 2014 [22]	inpainting (TCP and ETS)	18 tampered x2	Only 4 sequences available for direct download
VISION 2017 [124]	source/social media platform identification	1914 sequences	648 straight-from-device videos, 622 YouTube and 644 WhatsApp. Future dataset extension planned via "MOSES" application [125]
Ardizzone and Mazzola 2015 [33]	copy-move	160 sequences	Sequences synthesised from [40] and CANTATA datasets
Newson et al 2014 [42]	inpainting	3 tampered	Masks supplied for two sequences, demonstration of inpainting rather than explicit tampering detection dataset
Le et al 2017 [1]	inpainting	53 sequences	Both object removal and object reconstruction, demonstration of inpainting rather than explicit tampering detection dataset

post-tampering. Any recompression during distribution allows compression errors to creep in and increases the difficulty of extracting a bit-accurate tamper mask (as in Figure 6). Moreover, some pixels may be part of the tampered region but remain unchanged in value, and this makes for noisy masks in need of post processing.

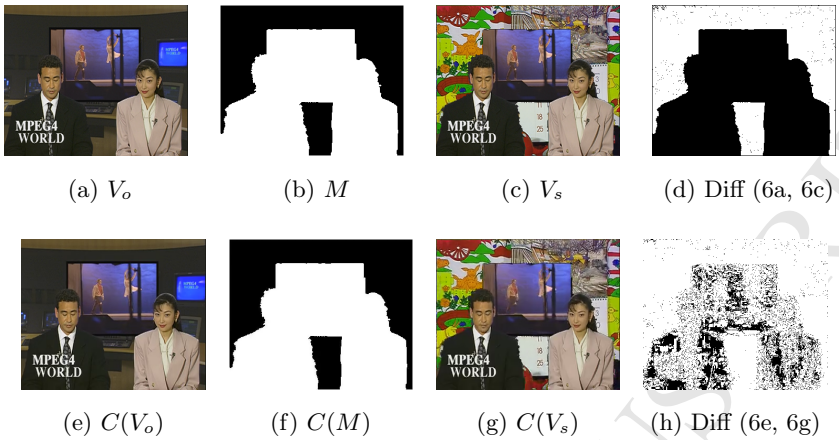
Using differences between original and tampered videos may be inappropriate for temporally tampered videos [98], where a frame-by-frame label might provide more information. This can be achieved when unprocessed original and tampered sequences are provided. Indeed, public inter-frame tampered video datasets are in short supply, with many inter-frame tampering detectors simply

building their own datasets from available sequences (see Section 2.2).

As can be seen in Table 3, most tampered video datasets focus on a single tampering method, such as splicing or object forgery or inpainting. Only VTD [98] demonstrates a variety of types. Variety is vital to accurately assess performance of video tampering detectors and support work towards universal video tampering detection. As discussed in [112], an approach using a combination of datasets will ensure more generalisable results with little need for specific domain adaptation. It is also important that tampered sequences are independent of tampering detection, so techniques such as [1, 33, 42] which publicly release their results are important to move forward both tampering AND tampering detection.

A number of datasets are produced and benchmarked with an existing detection technique [8, 22, 124] and many achieve high levels of precision on their selected dataset, often over 90% accuracy. This tends to portray tampering detection as a solved problem on that particular dataset, which discourages researchers from publishing lower results. A tampering detection method based on motion residue was presented in [126], and the experimental dataset was gathered from several previous object forgery works [23, 39, 127]. Accuracy was lower than 90%. This contrasts with over 90% in [127] and over 99% benchmarked in uncompressed FaceForensics [8]. However, in a real world situation, the original method of tampering will be unknown, therefore, it is worth collating results on several different datasets such that work towards a universal tampering detection can be realised.

Methods of dataset dissemination are an important consideration as this can cause unintentional post-processing of video data. Although video sharing websites such as YouTube may seem like an attractive distribution option, [98], any processing applied during publishing must be taken into account. It is possible to apply social media platform processing by uploading/downloading a video to/from a social media website, however the effects on the video are then irreversible. While researchers may add processing to an unprocessed video, they cannot remove it. Indeed, the effects of video processing by social media



(a) V_o (b) M (c) V_s (d) Diff (6a, 6c)

(e) $C(V_o)$ (f) $C(M)$ (g) $C(V_s)$ (h) Diff (6e, 6g)

Figure 6: The problems with recompression in the distribution of tampered datasets: Left column shows uncompressed data, right column has been lightly compressed. Figs 6e, 6f, 6g are the compressed versions of 6a, 6b, 6c respectively. Figs 6d and 6h are both uncompressed and show binarised differences.

platforms on video are represented in isolation in a dataset provided by [124] who found that sensor noise pattern used for camera source identification was adversely affected by processing on FaceBook, YouTube and WhatsApp, even when using high quality settings. These results are important in themselves as they show how tampering detection methods which rely on sensor noise, such as [26], can be defeated by virtue of the distribution platform alone. They also emphasise how post-processing can be easily overlooked.

Figure 6 illustrates some of the complications associated with recompression. Starting with uncompressed data, a binary mask was created based on segmentation of static and non-static content, and two uncompressed sequences simply spliced together. Figure 6d shows which pixels differ between Figures 6a and 6c and it can be seen that it is almost the perfect inverse of the mask (Fig 6b), with a few pixels that are identical between the original and spliced content. Figs 6e, 6f, 6g show the visual effects of compression on Figs 6a, 6b, 6c respectively. Figure 6h shows how compression has introduced tiny inaccuracies between the pixels of the original and spliced sequences so that the difference between them

no longer provides a mostly accurate inverse of the mask. With some thresholding and morphological processing, the difference sequence could still be used to infer a mask, but the degree of accuracy suffers even under slight compression. A compressed mask, as shown in Fig 6f provides a more accurate ground truth than deriving the mask from the compressed tampered/untampered pair. Moreover, official mask provision rather than frame difference inference removes the philosophical debate over whether a pixel, fully within the tampered region but by chance unchanged by the tampering process, is labelled tampered or not.

5. Conclusion

Many modern techniques of video tampering simply do not fit neatly into the traditional categories of inter- and intra- frame tampering. In particular, there is significant overlap between the recent categories of motion/style transfer and synthetic video generation. Changing the style of a video sequence from semantic segmentation masks to photo-realistic generates a purely synthetic video, but the same techniques can be used to perform digital puppetry and transfer motion from one mouth to another. This means that detection of synthetic video should be viewed as an extension of tampering detection. Given the current trend of using full reference quality measures in the evaluation of retouching, frame interpolation and in video frame prediction, it is clear that one of the current goals is to replicate authentic video. What remains unclear, however, is whether these methods will deviate from authentic content as evaluation methods emerge or even help to launder video tampering evidence in the same way as video compression.

One important new research direction in digital video manipulation is an accepted method of evaluation. Many existing methods rely on only qualitative evaluation and while this is an important first step, adoption of existing video quality techniques, including no reference quality metrics will speed up development. Until then, user studies and public release of manipulated video clips remains the gold standard. In the absence of elegant quality measures, altered

video and the associated methods are often publicly released for analysis, and video tampering detectors should look to utilise this provision where possible to create realistic detection methods. To facilitate this, video tamperers should release either sufficient data to simplify the creation of accurate tampering masks or release the masks themselves. Furthermore, video data should be distributed in such a way as to minimise further processing. Video processing, such as compression and retouching can effectively conceal tampering. While detection of such anti-forensics is an important research direction, processing can be applied to video independently after dataset publication, but only if the original dataset is published in such a way as to avoid unnecessary processing. With the increasing application of deep learning methods to tampering detection, any future dataset gatherers must take care to avoid potential pitfalls which cause datasets to reflect their own specific features relating to publishing platform or tool use, rather than those legitimately tied to the tampering technique.

As the variety of video manipulation techniques expands and advances, tampered and synthetic video will become indistinguishable from authentic video to human eyes. Therefore, new techniques are required which either classify tampered video according to its tampering type or perform tampering detection irrespective of the type of tampering. To maintain confidence in the authenticity of video content in future, it is crucial to develop techniques which can identify and localise video processing and manipulation. Universal video manipulation detection and localisation is essential if tampering detection is to keep pace with tampering methods.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] T. Le, A. Almansa, Y. Gousseau, S. Masnou, Motion-consistent video inpainting, in: ICIP 2017: IEEE International Conference on Image Processing, 2017.

[2] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, B. Catanzaro, Video-to-video synthesis, in: Advances in Neural Information Processing Systems, 2018.

[3] A. Bansal, S. Ma, D. Ramanan, Y. Sheikh, Recycle-gan: Unsupervised video retargeting, in: European Conference on Computer Vision, Springer, 2018, pp. 122–138.

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

[5] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems, 2017, pp. 700–708.

[6] H. Dong, P. Neekhara, C. Wu, Y. Guo, Unsupervised image-to-image translation with generative adversarial networks, arXiv preprint arXiv:1701.02676.

[7] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos, in: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, IEEE, 2016, pp. 2387–2395.

[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics: A large-scale video dataset for forgery detection in human faces, arXiv preprint arXiv:1803.09179.

[9] T. Karras, T. Aila, S. Laine, A. Herva, J. Lehtinen, Audio-driven facial animation by joint end-to-end learning of pose and emotion, ACM Transactions on Graphics (TOG) 36 (4) (2017) 94.

[10] L. Chen, Z. Li, R. K Maddox, Z. Duan, C. Xu, Lip movements generation at a glance, in: The European Conference on Computer Vision (ECCV), 2018.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[12] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, Youtube-8m: A large-scale video classification benchmark, arXiv preprint arXiv:1609.08675.

[13] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.

[14] K. Sitara, B. M. Mehtre, Digital video tampering detection: An overview of passive techniques, Digital Investigation 18 (2016) 8–22.

[15] R. D. Singh, N. Aggarwal, Video content authentication techniques: a comprehensive survey, Multimedia Systems (2017) 1–30.

[16] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, S. Tubaro, An overview on video forensics, APSIPA Transactions on Signal and Information Processing 1.

[17] R. C. Pandey, S. K. Singh, K. K. Shukla, Passive forensics in image and video using noise features: A review, Digital Investigation 19 (2016) 1–28.

[18] O. I. Al-Sanjary, G. Sulong, Detection of video forgery: A review of literature., Journal of Theoretical & Applied Information Technology 74 (2).

[19] A. Khodabakhsh, C. Busch, R. Ramachandra, A taxonomy of audiovisual fake multimedia content creation technology, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2018.

[20] Y. Wu, X. Jiang, T. Sun, W. Wang, Exposing video inter-frame forgery based on velocity field consistency, in: Acoustics, speech and signal processing (ICASSP), 2014 IEEE International Conference on, IEEE, 2014, pp. 2674–2678.

[21] K. Sitara, B. Mehtre, A comprehensive approach for exposing inter-frame video forgeries, in: Signal Processing & its Applications (CSPA), 2017 IEEE 13th International Colloquium on, IEEE, 2017, pp. 73–78.

[22] C.-S. Lin, J.-J. Tsay, A passive approach for effective detection and localization of region-level video forgery with spatio-temporal coherence analysis, Digital Investigation 11 (2) (2014) 120–140.

[23] K. A. Patwardhan, G. Sapiro, M. Bertalmío, Video inpainting under constrained camera motion, IEEE Transactions on Image Processing 16 (2) (2007) 545–553.

[24] A. Criminisi, P. Pérez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, IEEE Transactions on Image Processing 13 (9) (2004) 1200–1212.

[25] W. Wang, H. Farid, Exposing digital forgeries in interlaced and deinterlaced video, IEEE Transactions on Information Forensics and Security 2 (3) (2007) 438–449.

[26] M. Kobayashi, T. Okabe, Y. Sato, Detecting video forgeries based on noise characteristics, in: Pacific-Rim Symposium on Image and Video Technology, Springer, 2009, pp. 306–317.

[27] W.-H. Chuang, H. Su, M. Wu, Exploring compression effects for improved source camera identification using strongly compressed video, in: Image Processing (ICIP), 2011 18th IEEE International Conference on, IEEE, 2011, pp. 1953–1956.

[28] A. Subramanyam, S. Emmanuel, Video forgery detection using hog features and compression properties, in: Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on, IEEE, 2012, pp. 89–94.

[29] J. Kaur, S. Upadhyay, A. Sharma, A video database for intelligent video authentication, in: Computing, Communication and Automation (ICCCA), 2017 International Conference on, IEEE, 2017, pp. 1081–1085.

[30] J. A. Aghamaleki, A. Behrad, Malicious inter-frame video tampering detection in mpeg videos using time and spatial domain analysis of quantization effects, Multimedia Tools and Applications 76 (20) (2017) 20691–20717.

[31] M. A. Qureshi, M. Deriche, A bibliography of pixel-based blind image forgery detection techniques, Signal Processing: Image Communication 39 (2015) 46–74.

[32] V. Conotter, J. F. O'Brien, H. Farid, Exposing digital forgeries in ballistic motion, IEEE Transactions on Information Forensics and Security 7 (1) (2012) 283–296.

[33] E. Ardizzone, G. Mazzola, A tool to support the creation of datasets of tampered videos, in: International Conference on Image Analysis and Processing, Springer, 2015, pp. 665–675.

[34] D. D'Avino, D. Cozzolino, G. Poggi, L. Verdoliva, Autoencoder with recurrent neural networks for video forgery detection, Electronic Imaging 2017 (7) (2017) 92–99.

[35] C. Chan, S. Ginosar, T. Zhou, A. A. Efros, Everybody dance now, arXiv preprint arXiv:1808.07371.

[36] M. A. Qureshi, M. Deriche, A. Beghdadi, A. Amin, A critical survey of state-of-the-art image inpainting quality assessment metrics, Journal of Visual Communication and Image Representation 49 (2017) 177–191.

[37] T. Ha, S. Lee, J. Kim, Motion compensated frame interpolation by new block-based motion estimation algorithm, IEEE Transactions on Consumer Electronics 50 (2) (2004) 752–759.

[38] Y. Wexler, E. Shechtman, M. Irani, Space-time completion of video, IEEE Transactions on Pattern Analysis & Machine Intelligence (3) (2007) 463–476.

[39] T. K. Shih, N. C. Tang, J. C. Tsai, J.-N. Hwang, Video motion interpolation for special effect applications, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 41 (5) (2011) 720–732.

[40] G. Qadir, S. Yahaya, A. T. Ho, Surrey university library for forensic analysis (sulfa) of video content.

[41] P. Bestagini, S. Milani, M. Tagliasacchi, S. Tubaro, Local tampering detection in video sequences, in: Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, IEEE, 2013, pp. 488–493.

[42] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, P. Pérez, Video inpainting of complex scenes, SIAM Journal on Imaging Sciences 7 (4) (2014) 1993–2019.

[43] M. Ebdelli, O. Le Meur, C. Guillemot, Video inpainting with short-term windows: application to object removal and error concealment, IEEE Transactions on Image Processing 24 (10) (2015) 3034–3047.

[44] W. Lotter, G. Kreiman, D. Cox, Unsupervised learning of visual structure using predictive generative networks, arXiv preprint arXiv:1511.06380.

[45] Y. Dar, A. M. Bruckstein, Motion-compensated coding and frame rate upconversion: Models and analysis, IEEE Transactions on Image Processing 24 (7) (2015) 2051–2066.

[46] S. Niklaus, L. Mai, F. Liu, Video frame interpolation via adaptive separable convolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 261–270.

[47] S. Tulyakov, M.-Y. Liu, X. Yang, J. Kautz, Mocogan: Decomposing motion and content for video generation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[48] J. Walker, K. Marino, A. Gupta, M. Hebert, The pose knows: Video forecasting by generating pose futures, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 3352–3361.

[49] M. Jin, G. Meishvili, P. Favaro, Learning to extract a video sequence from a single motion-blurred image, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[50] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, N. Sebe, Every smile is unique: Landmark-guided diverse smile generation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[51] W. Xiong, W. Luo, L. Ma, W. Liu, J. Luo, Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[52] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, S. Levine, Stochastic variational video prediction, arXiv preprint arXiv:1710.11252.

[53] L. Zhao, X. Peng, Y. Tian, M. Kapadia, D. Metaxas, Learning to forecast and refine residual motion for image-to-video generation, in: The European Conference on Computer Vision (ECCV), 2018.

[54] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, D. Lin, Pose guided human video generation, in: The European Conference on Computer Vision (ECCV), 2018.

[55] F. A. Reda, G. Liu, K. J. Shih, R. Kirby, J. Barker, D. Tarjan, A. Tao, B. Catanzaro, Sdc-net: Video prediction using spatially-displaced convolution, in: The European Conference on Computer Vision (ECCV), 2018.

[56] H. Cai, C. Bai, Y.-W. Tai, C.-K. Tang, Deep video generation, prediction and completion of human action sequences, in: The European Conference on Computer Vision (ECCV), 2018.

[57] C. Cotsaces, N. Nikolaidis, I. Pitas, Video shot boundary detection and condensed representation: a review, IEEE Signal Processing Magazine 23 (2) (2006) 28–37.

[58] C. C. Huang, Y. Zhang, V. L. Thing, Inter-frame video forgery detection based on multi-level subtraction approach for realistic video forensic applications, in: Signal and Image Processing (ICSIP), 2017 IEEE 2nd International Conference on, IEEE, 2017, pp. 20–24.

[59] L. Zheng, T. Sun, Y.-Q. Shi, Inter-frame video forgery detection based on block-wise brightness variance descriptor, in: International Workshop on Digital Watermarking, Springer, 2014, pp. 18–30.

[60] E. Smith, A. Basharat, C. Anthony Hoogs, et al., A c3d-based convolutional neural network for frame dropping detection in a single video shot, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 86–94.

[61] D. Tralic, S. Grgic, B. Zovko-Cihlar, Video frame copy-move forgery detection based on cellular automata and local binary patterns, in: Telecommunications (BIHTEL), 2014 X International Symposium on, IEEE, 2014, pp. 1–4.

[62] M. C. Stamm, W. S. Lin, K. R. Liu, Temporal forensics and anti-forensics for motion compensated video, IEEE Transactions on Information Forensics and Security 7 (4) (2012) 1315–1329.

[63] Xiph.org video test media [derf's collection].
URL `https://media.xiph.org/video/derf/`

[64] M. K. Thakur, V. Saxena, J. Gupta, Learning based no reference algorithm for dropped frame identification in uncompressed video, in: Infor-

mation Systems Design and Intelligent Applications, Springer, 2016, pp. 451–459.

[65] L. He, W. Lu, C. Jia, L. Hao, Video quality assessment by compact representation of energy in 3d-dct domain, Neurocomputing 269 (2017) 108–116.

[66] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, M.-H. Yang, Learning blind video temporal consistency, in: The European Conference on Computer Vision (ECCV), 2018.

[67] V. Joshi, S. Jain, Tampering detection in digital video-a review of temporal fingerprints based techniques, in: Computing for sustainable global development (INDIACom), 2015 2nd International Conference on, IEEE, 2015, pp. 1121–1124.

[68] ITU-T, H.262 Information technology - Generic coding of moving pictures and associated audio information: Video, ITU-T (2 2012).

[69] ITU-T, H.264 Advanced video coding for generic audiovisual services, ITU-T (10 2016).

[70] H.-Y. Choi, H.-U. Jang, D. Kim, J. Son, S.-M. Mun, S. Choi, H.-K. Lee, Detecting composite image manipulation based on deep neural networks, in: Systems, Signals and Image Processing (IWSSIP), 2017 International Conference on, IEEE, 2017, pp. 1–5.

[71] S. Chen, S. Tan, B. Li, J. Huang, Automatic detection of object-based forgery in advanced video, IEEE Transactions on Circuits and Systems for Video Technology 26 (11) (2016) 2138–2151.

[72] P. Johnston, E. Elyan, C. Jayne, Spatial effects of video compression on classification in convolutional neural networks., in: Neural Networks (IJCNN), 2018 International Joint Conference on, IEEE, 2018, pp. 1370–1377.

[73] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[74] B. Bayar, M. 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, ACM, 2016, pp. 5–10.

[75] D. Cozzolino, G. Poggi, L. Verdoliva, Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection, in: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, ACM, 2017, pp. 159–164.

[76] C. Dong, Y. Deng, C. Change Loy, X. Tang, Compression artifacts reduction by a deep convolutional network, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 576–584.

[77] L. Cavigelli, P. Hager, L. Benini, Cas-cnn: A deep convolutional neural network for image compression artifact suppression, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE, 2017, pp. 752–759.

[78] J. Guo, H. Chao, One-to-many network for visually pleasing compression artifacts reduction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3038–3047.

[79] O. Kirmemis, G. Bakar, A. Murat Tekalp, Learned compression artifact removal by deep residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2602–2605.

[80] ITU-T, H.265 High efficiency video coding, ITU-T (12 2016).

[81] R. Yang, M. Xu, Z. Wang, T. Li, Multi-frame quality enhancement for compressed video, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[82] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (4) (2004) 600–612.

[83] A. Gironi, M. Fontani, T. Bianchi, A. Piva, M. Barni, A video forensic technique for detecting frame deletion and insertion., in: ICASSP, 2014, pp. 6226–6230.

[84] M. S. M. Sajjadi, R. Vemulapalli, M. Brown, Frame-recurrent video super-resolution, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[85] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, J. Kautz, Super slomo: High quality estimation of multiple intermediate frames for video interpolation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[86] M. Xia, G. Yang, L. Li, R. Li, X. Sun, Detecting video frame rate up-conversion based on frame-level analysis of average texture variation, Multimedia Tools and Applications 76 (6) (2017) 8399–8421.

[87] R. Li, Z. Liu, Y. Zhang, Y. Li, Z. Fu, Noise-level estimation based detection of motion-compensated frame interpolation in video sequences, Multimedia Tools and Applications 77 (1) (2018) 663–688.

[88] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, W. Shi, Real-time video super-resolution with spatio-temporal networks and motion compensation, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, 2017, pp. 2848–2857.

[89] R. Liao, X. Tao, R. Li, Z. Ma, J. Jia, Video super-resolution via deep draft-ensemble learning, in: The IEEE International Conference on Computer Vision (ICCV), 2015.

[90] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, J. Matas, Deblurgan: Blind motion deblurring using conditional adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[91] S. Nah, T. Hyun Kim, K. Mu Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3883–3891.

[92] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross, C. Schroers, Phasenet for video frame interpolation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[93] J. Janai, F. Güney, J. Wulff, M. J. Black, A. Geiger, Slow flow: Exploiting high-speed cameras for accurate and diverse optical flow reference data, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, 2017, pp. 1406–1416.

[94] S. Ilan, A. Shamir, A survey on data-driven video completion, in: Computer Graphics Forum, Vol. 34, Wiley Online Library, 2015, pp. 60–85.

[95] J. Chen, X. Kang, Y. Liu, Z. J. Wang, Median filtering forensics based on convolutional neural networks, IEEE Signal Processing Letters 22 (11) (2015) 1849–1853.

[96] D. Chauhan, D. Kasat, S. Jain, V. Thakare, Survey on keypoint based copy-move forgery detection methods on image, Procedia Computer Science 85 (2016) 206–212.

[97] J. Li, X. Li, B. Yang, X. Sun, Segmentation-based image copy-move forgery detection scheme, IEEE Transactions on Information Forensics and Security 10 (3) (2015) 507–518.

[98] O. I. Al-Sanjary, A. A. Ahmed, G. Sulong, Development of a video tampering dataset for forensic investigation, Forensic Science International 266 (2016) 565–572.

[99] Y. Aksoy, T. O. Aydin, M. Pollefeys, A. Smolić, Interactive high-quality green-screen keying via color unmixing, ACM Transactions on Graphics (TOG) 35 (5) (2016) 152.

[100] P. Bideau, E. Learned-Miller, Its moving! a probabilistic model for causal motion segmentation in moving camera videos, in: European Conference on Computer Vision, Springer, 2016, pp. 433–449.

[101] C. Xu, J. J. Corso, Libsvx: A supervoxel library and benchmark for early video processing, International Journal of Computer Vision 119 (3) (2016) 272–290.

[102] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[103] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784.

[104] P. Isola, J. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5967–5976. doi:10.1109/CVPR.2017.632.

[105] J. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251. doi:10.1109/ICCV.2017.244.

[106] F. Luan, S. Paris, E. Shechtman, K. Bala, Deep painterly harmonization, arXiv preprint arXiv:1804.03189.

[107] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (7) (2014) 1325–1339.

[108] J. Xu, B. Ni, Z. Li, S. Cheng, X. Yang, Structure preserving video prediction, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[109] J. He, A. Lehrmann, J. Marino, G. Mori, L. Sigal, Probabilistic video generation using holistic attribute control, in: The European Conference on Computer Vision (ECCV), 2018.

[110] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, C. Schmid, Learning from synthetic humans, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), IEEE, 2017, pp. 4627–4635.

[111] S. R. Richter, Z. Hayder, V. Koltun, Playing for benchmarks, in: International Conference on Computer Vision (ICCV), Vol. 2, 2017.

[112] A. Torralba, A. A. Efros, Unbiased look at dataset bias, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1521–1528.

[113] Credits for the use of the casia image tempering detection evaluation database (casia tide) v2.0 are given to the national laboratory of pattern recognition, institute of automation, chinese academy of science, corel image database and the photographers. http://forensics.idealtest.org.

[114] Y. Rao, J. Ni, A deep learning approach to detection of splicing and copy-move forgeries in images, in: Information Forensics and Security (WIFS), 2016 IEEE International Workshop on, IEEE, 2016, pp. 1–6.

[115] P. Sutthiwan, Y. Q. Shi, H. Zhao, T.-T. Ng, W. Su, Markovian rake transform for digital image tampering detection, in: Transactions on data hiding and multimedia security VI, Springer, 2011, pp. 1–17.

[116] P. Rota, E. Sangineto, V. Conotter, C. Pramerdorfer, Bad teacher or unruly student: Can deep learning say something in image forensics analysis?, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, 2016, pp. 2503–2508.

[117] S. Agarwal, W. Fan, H. Farid, A diverse large-scale dataset for evaluating rebroadcast attacks, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2018.

[118] H. Cao, A. C. Kot, Identification of recaptured photographs on lcd screens, in: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, 2010, pp. 1790–1793.

[119] P. Bas, T. Filler, T. Pevnỳ, Break our steganographic system: The ins and outs of organizing boss, in: International Workshop on Information Hiding, Springer, 2011, pp. 59–70.

[120] T. Gloe, R. Böhme, The dresden image database for benchmarking digital image forensics, Journal of Digital Forensic Practice 3 (2-4) (2010) 150–159.

[121] I. Amerini, T. Uricchio, L. Ballan, R. Caldelli, Localization of jpeg double compression through multi-domain convolutional neural networks, in: Proc. of IEEE CVPR Workshop on Media Forensics, Vol. 3, 2017.

[122] G. Schaefer, M. Stich, Ucid: An uncompressed color image database, in: Storage and Retrieval Methods and Applications for Multimedia 2004, Vol. 5307, International Society for Optics and Photonics, 2003, pp. 472–481.

[123] M. Boroumand, J. Fridrich, Deep learning for detecting processing history of images, Electronic Imaging 2018 (7) (2018) 1–9.

[124] D. Shullani, M. Fontani, M. Iuliani, O. Al Shaya, A. Piva, Vision: a video and image dataset for source identification, EURASIP Journal on Information Security 2017 (1) (2017) 15.

[125] D. Shullani, O. Al Shaya, M. Iuliani, M. Fontani, A. Piva, A dataset for forensic analysis of videos in the wild, in: International Tyrrhenian Workshop on Digital Communication, Springer, 2017, pp. 84–94.

[126] K. Kancherla, S. Mukkamala, Novel blind video forgery detection using markov models on motion residue, in: Asian Conference on Intelligent Information and Database Systems, Springer, 2012, pp. 308–315.

[127] C.-C. Hsu, T.-Y. Hung, C.-W. Lin, C.-T. Hsu, Video forgery detection using correlation of noise residue, in: Multimedia Signal Processing, 2008 IEEE 10th Workshop on, IEEE, 2008, pp. 170–174.