IEICE TRANS. INF. & SYST., VOL.E104–D, NO.1 JANUARY 2021

INVITED PAPER *Special Section on Enriched Multimedia — Multimedia Security and Forensics —*

Generation and Detection of Media Clones**

Isao ECHIZEN[†a)], **Noboru BABAGUCHI**[††], ***Fellows***, **Junichi YAMAGISHI**[†], **Naoko NITTA**[††], **Yuta NAKASHIMA**[††], **Kazuaki NAKAMURA**[††], **Kazuhiro KONO**[†††], ***Members***, **Fuming FANG**[†], ***Nonmember***, **Seiko MYOJIN**[††], ***Member***, **Zhenzhong KUANG**[††*], **Huy H. NGUYEN**[†], ***and*** **Ngoc-Dung T. TIEU**[†], ***Nonmembers***

SUMMARY With the spread of high-performance sensors and social network services (SNS) and the remarkable advances in machine learning technologies, fake media such as fake videos, spoofed voices, and fake reviews that are generated using high-quality learning data and are very close to the real thing are causing serious social problems. We launched a research project, the Media Clone (MC) project, to protect receivers of replicas of real media called media clones (MCs) skillfully fabricated by means of media processing technologies. Our aim is to achieve a communication system that can defend against MC attacks and help ensure safe and reliable communication. This paper describes the results of research in two of the five themes in the MC project: 1) verification of the capability of generating various types of media clones such as audio, visual, and text derived from fake information and 2) realization of a protection shield for media clones' attacks by recognizing them.

key words: *media clone, information security, image and speech processing, social media, communication model*

1. Introduction

With the spread of high-performance sensors and social network services (SNS), vast amounts of biometric identity information such as faces, voice, and human motion and vast amounts of natural language sources such as comments and reviews are being shared in cyberspace. Thanks to progress in machine learning technologies in recent years, the use of this vast amount of information as learning data has facilitated the generation of non-real voices, facial images, human motion videos, and natural language texts that are virtually real. We call such media "media clones (MCs)."

The creation of MCs (MC generation) is now being beneficially applied to a wide variety of services in the entertainment and communication fields, but it is also being used for fraud and misrepresentation, which poses the threat of spoofing. As a countermeasure against this threat, MC detection has been attracting widespread attention, and many applications such as ones for detecting fake videos and spoofed voices are being developed. We launched a research project called "Media Clone" aimed at achieving a communication system that can defend against MC attacks [1]. The project has five themes: (A) development of methods for protecting privacy, biological, and environmental information to prevent fake information generation, (B) verification of the capability of generating various types of MCs such as audio, visual, and text derived from fake information, (C) realization of a protection shield for MC attacks by recognizing them, (D) modeling of the envisioned communication system, and (E) experimental evaluation of the complete system and its components. We intend to build and open to the public a benchmark database of both real media and MCs. This paper describes the results of research in themes B and C. The results of themes A and D are described elsewhere [2].

Manuscript received April 7, 2020.
Manuscript revised July 12, 2020.
Manuscript publicized October 19, 2020.

[†]The authors are with National Institute of Informatics, Tokyo, 101–8430 Japan.
[††]The authors are with Osaka University, Suita-shi, 565–0871 Japan.
[†††]The author is with Kansai University, Takatsuki-shi, 569–1098 Japan.
[*]Presently, with Hangzhou Dianzi University, Hangzhou, Zhejiang, China.

[**]This work was supported in part by JSPS KAKENHI Grant Numbers JP16H06302, JP18H04120, and JST CREST Grant Number JPMJCR18A6, Japan.

a) E-mail: iechizen@nii.ac.jp

DOI: 10.1587/transinf.2020MUI0002

2. Related Work

Remarkable advances have been made in both generating and detecting various MC types, including speech, face images/videos, full-body motion videos, and texts. We briefly introduce several of these advances.

2.1 Media Clone Generation

Text-to-speech synthesis is a commonly used method for generating fake speech. Recent advances in deep neural networks have greatly enhanced the quality of fake speech. For example, the WaveNet autoregressive waveform model directly generates speech waveforms from linguistic or acoustic features [3], [4]. The DeepVoice [5] and Tacotron [6] models generate acoustic features directly from the input text. The Tacotron2 [7] model, which combines WaveNet with the Tacotron architecture, can perform end-to-end speech synthesis and generate human-like natural-sounding speech. Such models can be further adapted to separate speakers to achieve *voice cloning* [8]. Another approach for generating fake speech is *voice conversion* (VC), which transforms the speech of one person into the speech of another person. The VC framework has also greatly advanced. Recent conversion frameworks using neural waveform models can produce converted voice signals with greatly improved naturalness and speaker similarity [9].

Extensive research has been done on generating fake face images/videos [10]. *Face synthesis* is to generate photo-realistic non-existent face images. Generative adversarial networks (GANs) such as StyleGAN [11] are often used for generating high-quality fake face images. Face images can also be manipulated by using conventional computer graphics-based techniques such as FaceSwap [12] or by using deep-learning-based techniques such as autoencoder-based Deepfake [13]. Further, *facial reenactment* is to change the person's facial expressions in the given video to generate a fake video of that person [14] and can be realized by transferring the facial expression of one person to another person. Speech can also be used to generate fake videos by changing the facial expression of the person in the video [15]. While there has been less work on generating fake full-body motion videos, methods for facial reenactment have recently been applied to transfer the movements of one person to another person [16].

Deep neural networks have also been used extensively to generate fake texts. Since text is a sequence of words and/or characters, recurrent neural networks (RNNs), which can predict a probability distribution over the next element of a variable-length sequence, can be used to generate texts [17]. Its variants such as long short-term memory (LSTM) and gated recurrent unit (GRU) have also been used to improve the quality of generated texts [18]. While these networks usually focus on the content of texts, they can also be used to generate handwriting texts by predicting the pen's next coordinates in the handwriting strokes [19].

2.2 Media Clone Detection

Many fake detection methods have been proposed, especially for face images. The datasets play crucial roles in both training detectors and evaluating their performance. FaceForensics++ [20] contains fake face videos generated using face-swap techniques such as FaceSwap [12] and DeepFake [13] and face-reenactment techniques such as Face2Face [21] and NeuralTextures [22]. Facebook and other organizations recently conducted the Deepfake Detection Challenge and released a preview dataset [23]. DeepForensics-1.0 [24] is a large-scale dataset containing fake face videos of more diversity and higher quality generated by face-swap techniques.

In the image domain, detection methods can be classified into three groups: (1) using available architectures, (2) proposing new network architectures with default or custom layers, and (3) combining (1) and (2). As examples of (1), Raghavendra et al. [25] fine-tuned two available CNNs while Rossler et al. [20] used only one CNN. As examples of (2), Rahmouni et al. [26], Afchar et al. [27], and Quan et al. [28] proposed their own networks.

Besides the image domain, several approaches utilize temporal information from videos and additional information from audio. Li et al.'s work [29] focused on detecting eye blinking. Agarwal et al. [30] modeled facial expressions and movements to distinguish between real and fake videos. Sabir et al. [31] proposed a method that combines a recurrent CNN with a face alignment approach to improve detection accuracy.

3. Media Clone Generation

Our proposals for media clone generation cover a wide range of modalities, including voice, face, human body motion, and text (review and SNS posts). The following shows our research results on media clone generation.

3.1 Text-to-Speech Synthesizer for Synthesizing Celebrity Voices Using Speech Data Taken from the Web [32]

As described above, the outputs of a state-of-the-art speech synthesizer were rated as natural as human speech [7]. However, only studio-recorded highly controlled speech data were mostly used for building the speech synthesis models. If we consider more realistic spoofing scenarios, it is reasonable to assume that attackers use speech data taken from somewhere (e.g., the Internet), so the audio files used were likely recorded in non-controlled environments. We therefore first studied the feasibility of training a neural network to remove background noise from such audio and then trained state-of-the-art text-to-speech (TTS) systems to determine whether synthetic speech sounds natural [32].

For this purpose, we used publicly available speech data for President Barack Obama (a common target for identity theft research [33], [34]) and built a generative adversarial network-based speech enhancement system called SEGAN [35] and a WaveNet-based text-to-speech synthesizer [36]. The experimental results demonstrated that the enhancement system greatly improved the signal-to-noise ratio of low-quality speech data taken from publicly available sources; however, the synthetic speech generated by the WaveNet synthesizer was perceptually different [32].

3.2 Multi-Speaker Text-to-Speech Synthesizer Using Noise-Robust Neural Speaker Encoder [37]

We subsequently improved our end-to-end speech synthesizer so it can reproduce speaker characteristics even for noisy audio signals [37]. The key ideas are to train a noise-robust speaker classifier using large amounts of speech data with various types of background noise and reverberation, extract a hidden layer in the speaker classifier as a neural speaker embedding vector, and train the encoder-decoder end-to-end speech synthesizer, as shown in Fig. 1, using a high-quality multi-speaker clean speech database without noise. The speaker embedding vector is used as additional conditional input to the decoder network to generate acoustic features. Speech waveforms are generated via WaveNet, which is also trained using a high-quality multi-speaker speech database.

We found that this framework is robust to even noisy low-quality audio signals because its speaker encoder was

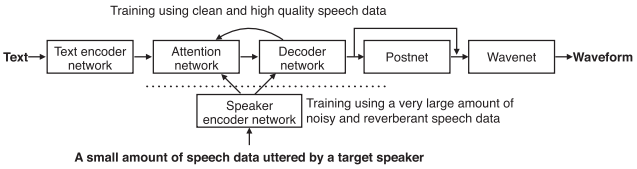


Fig. 1 Diagram of a noise robust multi-speaker TTS system proposed in [37]. Input is text, and output is speech waveforms. It consits of text encoder, attention, decoder, post-net, speaker encoder, and Wavenet blocks.

trained using large amounts of degraded speech data, enabling it to robustly estimate the speaker embedding vector. We also found that the output synthetic speech was of high quality [37] because the TTS network is trained using only a high-quality speech database separate from the speaker encoder network.

3.3 CycleGAN-Based Non-Parallel Voice Conversion [38]

Voice conversion (VC) is a speech processing technique by which the speech signals of a source speaker are modified to match those of the target speaker so that the modified speech sounds as if it was spoken by the target speaker. The main problem with VC is finding a mapping function between the two speakers that uses a corpus recorded by them. Depending on whether the training data are recorded by having the speakers speak the same sentences, the training data can be categorized as *parallel data* or *non-parallel data*. A VC system can be trained more accurately if parallel data are used since similar features of the two speakers can be easily aligned. The training is less accurate if non-parallel data are used because it is more difficult to align their features. Nevertheless, the use of non-parallel data has many benefits such as easier collection of the training data (even from the Internet), so a VC system trained using this type of data is more practical.

We proposed a non-parallel data-based VC method [38] that uses a CycleGAN [39] to match the two speakers' feature distributions instead of aligning their similar features. CycleGAN consists of two basic GANs [40] connected by two cycle consistency losses. Adversarial losses learn a mapping to match the two speakers' distributions, and cycle consistency losses keep the linguistic information unchanged. We carried out a large-scale listening test to compare the proposed VC method with two modern VC methods as baselines (a deep neural network method and a GAN-based method). We found that the proposed method achieved significantly better naturalness and speaker similarity than the two baselines. This is because the proposed method avoided feature alignment between the two speakers and did not introduce any alignment errors.

3.4 Speech-Driven Face Generation [41]

A classic approach for talking head or face generation relies on 3D reconstruction of the target face, modifying it in accordance with the target speech or target face [42]–[44].

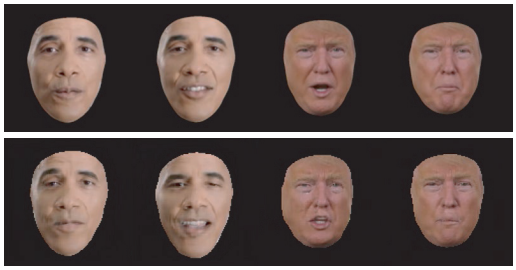


Fig. 2 Examples of ground-truth (top) and synthesized face images.

One of our proposed methods falls into this category [41]. More recent work took neural-network-based approaches, especially with the GAN framework [45].

We have also explored a neural-network-based approach in which the talking head of a person is synthesized using speech of that person obtained from a short (i.e., a few minutes) video clip of that person. A speech encoder maps the target speech into a representation correlated to mouth shape, which shares a similar goal to the speech recognition task. This is done by using as the encoder a latent representation computed at a certain layer of an off-the-shelf speech recognizer (e.g., [46]), enabling us to avoid training the encoder to model complex mapping from speech to the corresponding mouth shape. For better correlation with mouth shape, we train a mouth landmark encoder and then train a joint representation of a set of mouth landmarks and a speech segment. A generator, trained in a GAN framework, takes the learned representation of speech as input and generates a face image. For more consistent mouth shapes, we apply a mouth landmark regressor to the synthesized face image and use the mean squared error loss between the regressed and real landmark positions.

Figure 2 shows example synthesized facial images. For evaluation purposes, we used the audio track of a video so that the ground-truth face images could be extracted from the video. The synthesized faces are plausible; improving temporal consistency remains for future work.

3.5 Face Enhancement to Avoid Detection by Spoofing Detector [47]

A presentation attack is commonly used to bypass authentication systems using biometrics information (face, fingerprint, iris, and/or voice). Integration of a natural–CG image/video discriminator into the system before the authentication phase is one approach to preventing such attacks. To detect CG images in general, Wu et al. extracted statistical features from histograms of differential images [48]. In 2017, Peng et al. [49] reported a method based on multifractal and regression analysis. As an example of focusing on facial images, Nguyen et al.'s work [50] focused on facial smoothness as represented by edges and local entropy of the skin areas.

We developed a method for avoiding detection by those such detectors like those mentioned above [47]. It works

Table 1 Accuracy and detection rate before and after applying our proposed method on a custom database with images selected from MIT [52] and MS-Celeb database [53].

Spoofing detectors	Accuracy (%)		$\frac{TP}{TP+FN}$ (%)	
	Before	After	Before	After
Wu et al. [48]	56.38	**6.46**	100.00	**0.19**
Peng et al. [49]	92.32	**42.57**	100.00	**0.49**
Nguyen et al. [50]	96.72	**71.54**	99.20	**48.89**

by transforming CG images input to facial authentication systems to make them appear more natural. Our work is motivated by the idea of "adversarial machine learning" of Huang et al. for attacking machine learning based systems [51]. Although we did not target a specific system, we used a spoofing detection algorithm proposed by Wu et al. [48] as the basis of the discriminator. Given two sets of data (a set of natural images and a set of CG ones which are not necessarily corresponding person-to-person or pose-to-pose), a system using the proposed method implemented as a CNN transforms the CG input images in an attempt to make them indistinguishable from the natural counterparts.

An evaluation result were shown in Table 1. The transformed images again significantly reduced the detection rates of all spoofing detectors, especially those of Wu et al. [48] and Peng et al. [49], which were nearly 0%. Nguyen et al.'s method [50] had a detection rate of around 50%, down from nearly 100%, meaning that the attacker had a 50–50 chance of avoiding detection by this spoofing detector. Since the facial features were preserved, facial recognition was unaffected, therefore the transformed images could be used to bypass face authentication systems with a presentation attack detector integrated.

3.6 Audiovisual Face and Voice Transformation [54]

In previous subsections, we introduced VC and face generation independently. However, face movements and voice are correlated, especially lip movements and voice. This means that face movements and voice share common information, and we can use this information to improve face/voice transformation. Therefore we proposed a method for simultaneously transforming the face and voice of a source speaker into those of a target speaker by using a shared transformation neural network [54].

The first step in this method is to extract the mel-spectrogram from the speech data as the acoustic feature. The latent facial feature is extracted using the pre-trained VGG19 [55], and the facial keypoints are extracted using the OpenPose [56] as face geometric information. These features are then input into an audiovisual transformation network with stacked 1-D convolutional layers, and transform them to the target speaker's acoustic and facial features. The target speaker's facial image and speech are synthesized using an image reconstruction network and the WaveNet, respectively. Both the transformed acoustic and facial features are used as input for the image reconstruction network and WaveNet. During training, the described three networks were independently trained using parallel data (same utterances and facial expressions) recorded by the source and target speakers. During testing, the three networks were concatenated.

We compared the correlation scores between simultaneous and independent transformation and found that those for simultaneous transformation were higher. This suggests that transforming the facial and acoustic features together makes it possible for the transformed voice and facial expressions to be highly correlated. We also performed a human evaluation experiment to compare simultaneous with independent transformation in terms of naturalness and similarity. The simultaneous transformation again achieved better results. This suggests that acoustic and facial characteristics can compensate for each other and thus can be used to help transformation.

3.7 Motion Video Generation from a Still Image [57]

Techniques for generating human motion video are widely applicable to entertainment systems, sports coaching systems, video game production, and so on. We refer to automatically generated human motion videos as motion video clones (MVCs). MVCs are conventionally generated by constructing and manipulating a 3D human body model consisting of a textured mesh and a skeleton [58]. More recently, training-based approaches using GANs have been actively studied [22], [59], similar to the case of face cloning. However, it is labor-intensive and time-consuming to prepare a large amount of training data or a detailed 3D model of every possible system user. Hence, we proposed a model-free and training-free method [57].

The proposed method requires only two kinds of input data: a *reference video*, in which a person A performs some motion, and a *target image*, which is a single still image of another person B. Using them, our method generates an MVC in which person B virtually performs the same motion as person A. In this method, a 2D skeleton is first extracted from both the target image and the reference video, and the target's skeleton is manipulated so that it has the same posture as the reference's skeleton. Next, a 2D affine transform is computed between the original position and the manipulated position of the target's skeleton, which is separately done for each body part such as the left arm and the right leg. The computed affine transforms are used to map the texture of the target image onto each frame of the output MVC. This part-wise texture mapping makes the proposed method robust against self-occlusion of the body. However, it makes the areas around the body joints appear unnatural. To avoid this, we extend the part-wise affine transforms to pixel-wise ones. Specifically, we fuzzily cluster a set of pixels on the target image into the body parts, and the resultant membership values of each pixel are used to compute a linear combination of the part-wise affine transforms. An example of an MVC generated with the proposed method is shown in Fig. 3.

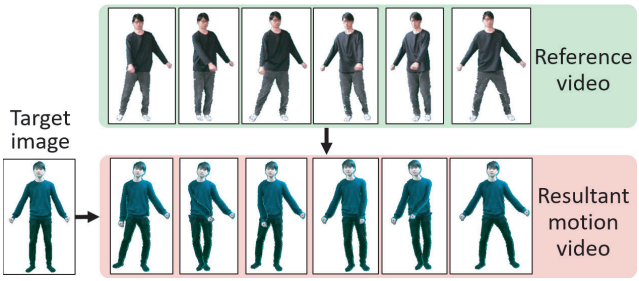


Fig. 3 Example motion video frames generated with model-free and training-free method [57].

Fig. 4 Actual handwritten characters and clones generated with proposed method [60].

3.8 Handwritten Character Image Generation [60]

There are several ways to represent text data: plain text, document images, character images, strokes, and so on. Among them, we focused on handwritten character images and proposed a method for generating clones [60]. The generated images, which we call handwritten character clones (HCCs), are useful for chat-like communication tools, assistance systems for hand-impaired people, and so on.

The goal with this method is to generate synthetic images that mimic a writer's actual handwritten texts character-by-character. It is generally difficult to collect a large training dataset of all characters written by a target writer, especially for languages having thousands of characters such as Chinese and Japanese. Hence, we assume that a single image can be used for some characters (*seed characters*) and that no image is available for the other characters (*non-seed characters*). To generate HCCs from such a limited training dataset, we use an auxiliary dataset containing characters written by many other writers. In the proposed method, a number of character shape models are first created for each character by clustering the character images in the auxiliary dataset. Next, for each seed character, the shape model that best fits the training data created by the target writer is chosen. Finally, for the non-seed characters, the best shape model for each one is estimated on the basis of collaborative filtering. In the generation stage, several HCCs are sampled from the chosen model for each character and concatenated in accordance with the given sentence. Figure 4 shows an example of actual handwritten characters and the clones generated with the proposed method.

3.9 Fake Review Generation [61]

In addition to the modalities mentioned above, fluent sentences and paragraphs can also be automatically generated using recently developed deep-learning-based neural language models. Applications of neural language models mainly include machine translation, image captioning, text summarization, dialogue generation, and speech recognition. Therefore, neural text generation has become an indispensable technique in the natural language processing field. Unfortunately, neural language models can also be used to generate text for use in fake news and fake reviews.

Here, we proposed a method for generating fake reviews that requires minimal skills but has high performance [61]. A user of this method can create a flood of positive or negative fake reviews to affect the rating of a product. A real review $\mathbf{x}$ is first selected from a shopping web site and input to the system. The GPT-2 language model [62] then uses it to automatically generate a huge number of fake reviews $\mathbf{x}'$. Because there are no constraints on review generation, the BERT [63] is then used to filter out reviews with undesired sentiments. This method works in combination with a language model, and many high-performance language models are publicly available. Since no special skills are needed to use this method, it poses a big risk. We carried out a human evaluation experiment using reviews taken from the Amazon and Yelp review corpora as seed reviews. We randomly displayed a real and three fake reviews on a web interface and asked 80 volunteers (39 native and 41 non-native English speakers) to evaluate the fluency of the reviews and identify the real review. The fake reviews had the same fluency score as the real reviews on average. The selection correctness was around 25%, meaning that the volunteers could not distinguish the fake reviews from the real ones.

3.10 Social Media Message Generation [64]

Spoofing in social networking services by impersonating legitimate users can be considered as a form of media cloning. With the recent development of deep-learning text generators, spoofing messages can be automatically generated if the target user has posted a sufficient number of messages to be used as the training data, and accounts posting such spoofing messages have been created, especially on Twitter. We further examined whether the spoofing messages can be generated even for target users who have posted only a limited number of messages [64]. Our idea is to collect messages reflecting the personality of each user, while also reflecting patterns shared among different users, as the training data. Since legitimate users exist in the real world, they often post about their experiences at various points of interest (PoIs), i.e., places or events that they find interesting. Furthermore, different users tend to post both semantically and syntactically similar messages for semantically similar PoIs. For example, messages about different

Table 2 Examples of generated messages

PoI	generated message
yankee stadium	go oriole! ♯bowlegesco ♯yankeestadium ♯soccer ♯love ♯friends
	take the final game to the yankee stadium. ♯pennantrace ♯yank ♯yankees ♯love
sea food city	fresh chicken and waffles. ♯cityisland ♯seafoodcity ♯foodporn ♯foodie
	eat a ♯cityland ♯vegan ♯chicken ♯seafood ♯foodporn
madison square	our favorite song of the year. ♯adeleconcert ♯adelelive2016 ♯madisonsquare ♯theadele
	take the stage with my favorite band. ♯ufc205 ♯madisonsquare ♯livemusic

restaurants tend to be about the food served. Since messages about PoIs at different locations typically contain words unique to each PoI, such as its names, they can be easily collected by examining the geospatial locality of words. The semantics of each PoI are then estimated from the collected messages, and a message generator is trained to generate messages in accordance with the semantics of a given PoI. In addition, the preferences of each target user can be estimated from even a small number of his/her previous messages, to automatically select a PoI that he/she might visit. When given the semantics of the selected PoI as the input, the trained message generator can generate a spoofing message that could be accepted as the message posted by the target user. Experiments demonstrated that the interests of the target user can be estimated from as few as 25 messages and that messages can be generated in accordance with the selected PoI. Examples of the generated messages are shown in Table 2.

4. Media Clone Detection

The detection of cloned media is relevant to anti-spoofing techniques for face recognition and anti-spoofing techniques for speaker recognition [65]. Anti-spoofing techniques are aimed at discriminating between fake artificial inputs presented to biometric authentication systems and genuine live inputs. Examples of fake inputs include 3D-printed facial mask attacks against face recognition systems and reply attacks against speaker recognition systems. If sufficient knowledge and data regarding the spoofed data is available, a binary classifier can be constructed for use in anti-spoofing systems. If sufficient knowledge and data are not available, one-class classifier techniques such as anomaly detection and outlier detection are used instead. Likewise, the detection of cloned media can also be defined as a binary discrimination task, i.e., discriminating between original genuine media and generated/manipulated media, or as a one-class classification task by assuming that generated/manipulated media are outliers. The following shows our research results on media clone detection.

Table 3 Accuracy of capsule network against several types of attack.

Database	Accuracy (%)
Replay Attack [69]	00.00
CGvsPhotos - Patches [26]	97.00
CGvsPhotos - Full size [26]	100.00
DeepFakes - Frame level [27]	95.93
DeepFakes - Video level [27]	99.23
FaceForensics - No compression [70]	99.37
FaceForensics - Light compression [70]	96.50
FaceForensics - Strong compression [70]	81.00

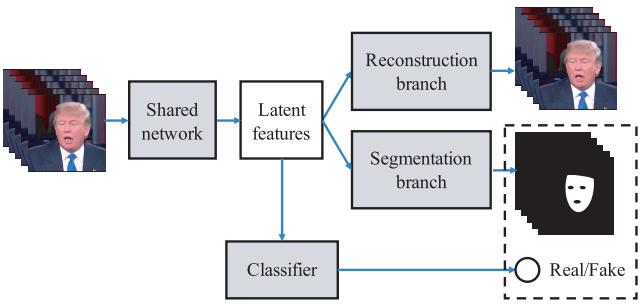


Fig. 5 Overview of the multitask network [67].

4.1 Detection of Computer Generated/Manipulated Faces [66], [67]

To detect computer generated/manipulated images, we proposed a high-performance classifier [66] based on the capsule network architecture [68], which has fewer parameters than traditional CNNs. The proposed capsule module consists of three primary capsules and two output capsules, one for real and one for fake images. Output from the primary capsules is routed to the output capsules using a dynamic routing algorithm. The proposed method can detect various kinds of fake facial images, including replay attacks using printed images or recorded videos and computer-generated/manipulated facial images created using computer-graphics-based approaches or CNN-based ones, which are summarized in Table 3.

Another important topic besides classification is locating manipulated regions. We introduced a Y-shape autoencoder network (Fig. 5) that uses multi-task learning and semi-supervised learning approaches to simultaneously detect manipulated images/video frames and locate the manipulated regions [67]. Activation of the encoded features is used for classification. One branch of the decoder is used for reconstructing the original input while the other branch is used for segmenting the manipulated regions. Information shared among the three tasks (classification, reconstruction, and segmentation) helps improve the overall performance of the network. Experiments on the FaceForensics [70] and FaceForensics++ [20] databases demonstrated the effectiveness of this Y-shape network and its generalizability under the mismatch condition for previously seen attacks. For unseen attacks, fine-tuning the network using only a small amount of new data is enough for the network to achieve

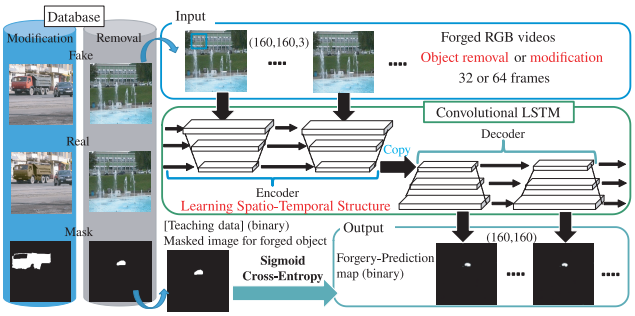


Fig. 6 Overview of passive video forgery detection system.

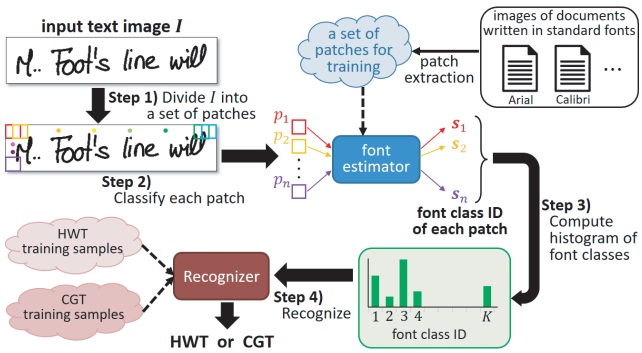


Fig. 7 Overview of method for discriminating computer-generated text from handwritten text.

good performance.

4.2 Passive Video Forgery Detection [71]

Video forgery detection is a major challenge in video processing. The wide variety of video and the various kinds of tampering make it hard for a forgery detection system to detect all forged videos. Most researchers thus impose restrictions on the kinds of videos and tampering, but this can make the system impractical. In our work, we treated dynamic videos, such as ones with dynamic backgrounds and changing perspectives. We also targeted several types of spatial tampering.

Another major challenge in this field is that there has been no work on object modification attacks. Spatial tampering is divided into three types of attack: object removal, addition, and modification [72]. Most researchers target either object addition or object removal attacks. We created a database for object modification that was created semi-automatically and adjusted every frame, and we addressed object modification and object removal attacks.

Our purpose was to achieve a passive video forgery detection system with an end-to-end architecture based on a deep neural network and with high detection accuracy. To take into account both the spatial and temporal consistencies of videos, we used a model based on a convolutional LSTM [71]. An overview of our proposed system is shown in Fig. 6. Testing demonstrated that both spatial and temporal information needs to be used in order to detect elaborately forged objects whereas roughly forged objects can be detected by using spatial information only.

4.3 Handwritten Text Image Detection [73]

Techniques for generating handwritten-like text images, such as the one mentioned in Sect. 3.8, can be maliciously exploited to forge documents. To detect computer-generated text (CGT) as clones, we proposed a method for discriminating CGT from images of handwriting text (HWT) written by a person [73]. In general, characters in HWT have various shapes and styles even when written by the same person. Although such variation is a characteristic property of HWT, it has not been quantitatively analyzed in previous work. Hence, even state-of-the-art CGT generation methods cannot perfectly mimic the variation in character shapes and styles in HWT. Focusing on this weakness, we developed a method that uses a distribution of patch-wise font features to discriminate between CGT and HWT. As shown in Fig. 7, the method first divides an input text image into a set of patches and classifies each patch into one of the pre-defined standard font classes. The font estimator is trained using images of digital documents. Then a histogram of the patch-wise font classes is computed and fed into a CGT/HWT recognizer as a feature vector. Testing showed that the proposed method could discriminate CGT generated using an RNN-based method [74] from HWT with an accuracy of 96.3%.

4.4 Fake Review Detection [61]

In Sect. 3.9, we introduced an easy-to-use but high-performance fake-review generation method. The fake reviews generated with this system can fool human readers. It is import to develop a method for automatically detecting them to help human readers identify fake reviews.

Several methods can be used to detect fake sentences, such as the Grover [75], the GLTR [76], and OpenAI GPT-2 detector [77]. The Grover is based on a neural network and was developed for detecting fake news, such as fake news generated using the GPT-2 language model. The GLTR was developed for detecting the tendencies of sentences. By comparing the tendencies of human-written sentences with those of fake sentences, we can say that a sentence appears to be real or fake. The OpenAI GPT-2 detector is a RoBERTa [78]-based language model that can be used as a text classifier. In addition, we can fuse these detectors by using logistic regression at the score level to further improve detection accuracy.

Table 4 shows the detection results [61]. We used the method described in Sect. 3.9 to generate fake reviews based on Amazon and Yelp reviews. We used the equal error rate (EER) as the metric. Among the detection methods, the GPT-2 detector achieved the best results. If we further fuse these detection methods using logistic regression at the score level, we can obtain a lower EER. However, the lowest error rate (19.6%) is not sufficient for practical use. This

Table 4 Equal error rate in distinguishing between fake and real reviews. (GPT-2D is OpenAI GPT-2 detector; "+" indicates score fusion.)

Detector	Amazon	Yelp	Overall
Grover	43.6%	36.9%	40.7%
GTLR	40.9%	35.9%	38.5%
GPT-2D	**20.9%**	25.8%	23.5%
Grover + GTLR	35.3%	34.6%	34.9%
Grover + GPT-2D	24.9%	22.2%	23.4%
GTLR + GPT-2D	25.0%	**19.6%**	**22.5%**
Grover + GTLR + GPT-2D	25.0%	**19.6%**	**22.5%**

means that further development of fake review detectors is needed.

5. Voice Conversion and Anti-Spoofing Challenges

In addition to advanced research on each modality, authors from National Institute of Informatics have conducted two challenges related to media cloning and its detection to provide a common evaluation platform and evaluation metrics for fair comparison: the Voice Conversion Challenge and the ASVspoof Challenge.

The Voice Conversion Challenge is a biannual event that started in 2016. In this challenge, we provide a common database to challenge participants, the participants build voice converters using their own technology, and the organizers rank converted speech provided by the participants. The main evaluation methodology is a listening test in which crowd-sourced participants evaluate the naturalness and speaker similarity of the converted speech. In the 2016 Challenge, a standard VC task using a parallel database was adopted [79]. The main focus of the 2018 Challenge was a more advanced conversion scenario using a non-parallel database (source and target speaker utterances differ) [32]. In the 2020 Challenge, the main task will be cross-lingual VC. All the speech converted by the participants have been released as open data, so researchers can compare the performance of their VC system with that of a state-of-the-art system without re-implementing it. All of the converted data are based on advanced VC systems, so they can be used by the biometric community to train and improve anti-spoofing models.

The spoofing capability against automatic speaker verification (ASV) has also been evaluated. The ASVspoof Challenge is also a biannual event. It started in 2013. Like in the Voice Conversion Challenge, we provide a common database including many pairs of spoofed audio (generated audio and/or replay audio) and genuine audio to challenge the participants, the participants build anti-spoofing models using their own technology, and the organizers rank the detection accuracy of the anti-spoofing models provided by the participants. The main metrics are the EER and a tandem decision cost function [80] that takes into account errors in both the ASV and anti-spoofing systems. In 2015, the first anti-spoofing database including various types of spoofed audio was constructed, and this database became the standard in the ASV community [81]. The main focus of the 2017 Challenge was a reply task, and a large quantity of real-world reply speech data was collected [82]. In 2019, an even larger database including both generated and reply speech data was constructed and distributed to over 150 organizations [83]. The results of this last Challenge revealed that although human listeners could not distinguish most advanced spoofed audio from genuine audio, the anti-spoofing systems could still differentiate them more accurately. This clearly demonstrates that human perception and machine perception are totally different, meaning they can complement each other.

6. Conclusion

We have outlined the major achievements of our Media Clone project aimed at the generation and detection of media clones. Such generation and detection are not only a problem of biometric authentication but also a very important issue from the viewpoint of the reliability of media such as videos and news reports. A recent example of this is the fake news about preventive and therapeutic methods related to the novel coronavirus infection (COVID-19) without scientific basis being spread on SNS, causing a flood of unreliable information, i.e., "a wave of infodemics," that caused fear and confusion in society. Attacker groups with specific intentions can thus generate media clones and use them to create a wave of infodemics on SNS. There is thus a risk that such groups will be able to easily induce thoughts and manipulate public opinion by repeatedly posting information that is not true. Future work thus includes investigating the reliability of information and social interaction through social media from the viewpoint of social science, which should lead to results with high social significance.

References

[1] "Communication system for defending against attacks of media clones." http://www2c.comm.eng.osaka-u.ac.jp/proj/mc/eindex.html, 2020.
[2] N. Babaguchi, I. Echizen, J. Yamagishi, N. Nitta, Y. Nakashima, K. Nakamura, K. Kono, F. Fang, S. Myojin, Z. Kuang, H.H. Nguyen, and N.D.T. Tieu, "Preventing fake information generation against media clone attacks," IEICE Trans. Inf. & Syst., vol.E104-D, no.1, pp.2–11, Jan. 2021.
[3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: a generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.
[4] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," Proc. Interspeech, pp.1118–1122, 2017.
[5] S.Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: real-time neural text-to-speech," Proc. International Conference on Machine Learning, pp.195–204, 2017.
[6] Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R.A. Saurous, "Tacotron: a fully end-to-end text-to-speech synthesis model," Proc. Interspeech, pp.4006–4010, 2017.
[7] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R.J. Skerry-Ryan, R.A. Saurous, Y. Agiomygiannakis, and Y. Wu, "Natural TTS synthesis by

conditioning WaveNet on mel spectrogram predictions," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4779–4783, 2018.

[8] S.Ö. Arik, J. Chen, K. Peng, W. Ping, and Z. Yanqi, "Neural voice cloning with a few samples," Advances in Neural Information Processing Systems, pp.10019–10029, 2018.

[9] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, pp.195–202, 2018.

[10] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: a survey of face manipulation and fake detection," arXiv preprint arXiv:2001.00179, 2020.

[11] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.4401–4410, 2019.

[12] "FaceSwap." https://github.com/MarekKowalski/FaceSwap.

[13] "Terrifying high-tech porn: Creepy 'deepfake' videos are on the rise." https://www.foxnews.com/tech/terrifying-high-tech-porn-creepy-deepfake-videos-are-on-the-rise. Accessed: 2018-02-17.

[14] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Matthias, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," ACM Transactions on Graphics, vol.37, no.4, pp.1–14, 2018.

[15] A. Jamaludin, J.S. Chung, and A. Zisserman, "You said that?: synthesising talking faces from audio," International Journal of Computer Vision, vol.127, no.11–12, pp.1767–1779, 2019.

[16] L. Liu, W. Xu, M. Zollhöfer, H. Kim, F. Bernard, M. Habermann, W. Wang, and C. Theobalt, "Neural rendering and reenactment of human actor videos," ACM Transactions on Graphics, vol.38, no.5, pp.1–14, 2019.

[17] I. Sutskever, J. Martens, and G. Hinton, "Generating text with recurrent neural networks," Proc. International Conference on Machine Learning, pp.1017–1024, 2011.

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[19] A. Graves, "Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2013.

[20] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: learning to detect manipulated facial images," Proc. International Conference on Computer Vision, pp.1–11, 2019.

[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[22] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," ACM Transactions on Graphics, vol.38, no.4, pp.1–12, 2019.

[23] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C.C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," arXiv preprint arXiv:1910.08854, 2019.

[24] L. Jiang, W. Wu, R. Li, C. Qian, and C.C. Loy, "DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection," arXiv preprint arXiv:2001.03024, 2020.

[25] R. Raghavendra, K.B. Raja, S. Venkatesh, and C. Busch, "Transferable deep-CNN features for detecting digital and print-scanned morphed face images," Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop, pp.1822–1830, 2017.

[26] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," Proc. IEEE International Workshop on Information Forensics and Security, pp.1–6, 2017.

[27] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a compact facial video forgery detection network," Proc. IEEE International Workshop on Information Forensics and Security, pp.1–7, 2018.

[28] W. Quan, K. Wang, D.-M. Yan, and X. Zhang, "Distinguishing between natural and computer-generated images using convolutional neural networks," IEEE Trans. Inf. Forensics Security, vol.13, no.11, pp.2772–2787, 2018.

[29] Y. Li, M.-C. Chang, H. Farid, and S. Lyu, "In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking," Proc. IEEE International Workshop on Information Forensics and Security, pp.1–7, 2018.

[30] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.38–45, 2019.

[31] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.80–87, 2019.

[32] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data," Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, pp.240–247, 2018.

[33] R. Kumar, J. Sotelo, K. Kumar, A. de Brebisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," NIPS 2017 Workshop on Machine Learning for Creativity and Design, 2017.

[34] S. Suwajanakorn, S.M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: learning lip sync from audio," ACM Transactions on Graphics, vol.36, no.4, pp.95:1–95:13, 2017.

[35] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," Proc. Interspeech, pp.3642–3646, 2017.

[36] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2018.

[37] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2020.

[38] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.5279–5283, 2018.

[39] J.-Y. Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Proc. International Conference on Computer Vision, pp.2223–2232, 2017.

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

[41] Y. Nakashima, T. Yasui, L. Nguyen, and N. Babaguchi, "Speech-driven face reenactment for a video sequence," ITE Transactions on Media Technology and Applications, vol.8, no.1, pp.60–68, 2020.

[42] Y. Liu, F. Xu, J. Chai, X. Tong, L. Wang, and Q. Huo, "Video-audio driven real-time facial animation," ACM Trans. Graphics, vol.34, no.6, pp.182:1–182:10, 2015.

[43] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," ACM Trans. Graphics, vol.37, no.4, pp.163:1–163:14, 2018.

[44] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Nießner, "Headon: Real-time reenactment of human portrait videos," ACM Transactions on Graphics, vol.37, no.4, pp.1–13, 2018.

[45] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," International Journal of Computer Vi-

sion, 2019.

[46] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "wav2letter++: The fastest open-source speech recognition system," arXiv preprint, arXiv:1812.07625, 2018.

[47] H.H. Nguyen, N.-D.T. Tieu, H.-Q. Nguyen-Son, J. Yamagishi, and I. Echizen, "Transformation on computer-generated facial image to avoid detection by spoofing detector," Proc. IEEE International Conference on Multimedia and Expo, pp.1–6, 2018.

[48] R. Wu, X. Li, and B. Yang, "Identifying computer generated graphics via histogram features," Proc. IEEE International Conference on Image Processing, pp.1933–1936, 2011.

[49] F. Peng, D.L. Zhou, M. Long, and X.M. Sun, "Discrimination of natural images and computer generated graphics based on multi-fractal and regression analysis," AEU-International Journal of Electronics and Communications, vol.71, pp.72–81, 2017.

[50] H.H. Nguyen, H.-Q. Nguyen-Son, T.D. Nguyen, and I. Echizen, "Discriminating between computer-generated facial images and natural ones using smoothness property and local entropy," Proc. International Workshop on Digital Watermarking, pp.39–50, 2015.

[51] L. Huang, A.D. Joseph, B. Nelson, B.I.P. Rubinstein, and J.D. Tygar, "Adversarial machine learning," Proc. ACM Workshop on Security and Artificial Intelligence, pp.43–58, 2011.

[52] B. Weyrauch, B. Heisele, J. Huang, and V. Blanz, "Component-based face recognition with 3d morphable models," Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop, p.85, 2004.

[53] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," Proc. European conference on computer vision, pp.87–102, 2016.

[54] F. Fang, X. Wang, J. Yamagishi, and I. Echizen, "Audiovisual speaker conversion: jointly and simultaneously transforming facial expression and acoustic characteristics," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.6795–6799, 2019.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[56] Z. Cao, G. Hidalgo, T. Simon, S.E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[57] T. Tsutumi, K. Nakamura, S. Myojin, N. Nitta, and N. Babaguchi, "Training-free method for generating motion video clones from a still image considering self-occlusion of human body," Proc. International Conference on Image Processing, pp.509–513, 2019.

[58] J.P. Lewis, M. Cordner, and N. Fong, "Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation," Proc. Annual Conference on Computer Graphics and Interactive Techniques, pp.165–172, 2000.

[59] C. Chan, S. Ginosar, T. Zhou, and A.A. Efros, "Everybody dance now," Proc. International Conference on Computer Vision, pp.5933–5942, 2019.

[60] K. Nakamura, E. Miyazaki, N. Nitta, and N. Babaguchi, "Generating handwritten character clones from an incomplete seed character set using collaborative filtering," Proc. International Conference on Frontiers in Handwriting Recognition, pp.68–73, 2018.

[61] D.I. Adelani, H. Mai, F. Fang, H.H. Nguyen, J. Yamagishi, and I. Echizen, "Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection," Proc. International Conference on Advanced Information Networking and Applications, pp.1–12, 2020.

[62] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, vol.1, no.8, p.9, 2019.

[63] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[64] J. Lim, N. Nitta, K. Nakamura, and N. Babaguchi, "Generating spoofing tweets considering points of interest of target user," Proc. APSIPA Annual Summit and Conference, pp.1672–1678, 2019.

[65] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," Speech Communication, vol.66, pp.130–153, 2015.

[66] H.H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp.2307–2311, 2019.

[67] H.H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," Proc. IEEE International Conference on Biometrics: Theory, Applications and System, 2019.

[68] S. Sabour, N. Frosst, and G.E. Hinton, "Dynamic routing between capsules," Proc. Conference on Neural Information Processing Systems, 2017.

[69] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," Proc. International Conference of the Biometrics Special Interest Group, 2012.

[70] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," arXiv preprint arXiv:1803.09179, 2018.

[71] K. Kono, T. Yoshida, S. Ohshiro, and N. Babaguchi, "Passive video forgery detection considering spatio-temporal consistency," Proc. International Conference on Soft Computing and Pattern Recognition, pp.381–391, 2018.

[72] N. Sowmya K and H. Chennamma, "A survey on video forgery detection," arXiv preprint arXiv:1503.00843, 2015.

[73] N. Hamasaki, K. Nakamura, N. Nitta, and N. Babaguchi, "Discrimination between handwritten and computer-generated texts using a distribution of patch-wise font features," Proc. APSIPA Annual Summit and Conference, pp.1665–1671, 2019.

[74] A. Graves, "Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2014.

[75] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," Advances in Neural Information Processing Systems, pp.9051–9062, 2019.

[76] S. Gehrmann, H. Strobelt, and A.M. Rush, "Gltr: Statistical detection and visualization of generated text," Proc. Annual Meeting of the Association for Computational Linguistics, 2019.

[77] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, and J. Wang, "Release strategies and the social impacts of language models," arXiv preprint arXiv:1908.09203, 2019.

[78] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[79] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," Proc. Interspeech, pp.1632–1636, 2016.

[80] T. Kinnunen, K.A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D.A. Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, pp.312–319, 2018.

[81] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," Proc. Interspeech, pp.2037–2041, 2015.

[82] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K.A. Lee, "The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection," Proc. Interspeech, pp.2–6, 2017.

[83] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T.H. Kinnunen, and K.A. Lee,

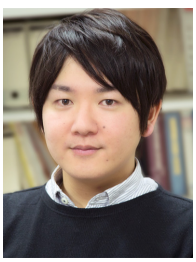
"ASVspoof 2019: future horizons in spoofed and fake audio detection," Proc. Interspeech, pp.1008–1012, 2019.

Isao Echizen received B.S., M.S., and D.E. degrees from the Tokyo Institute of Technology, Japan, in 1995, 1997, and 2003, respectively. He joined Hitachi, Ltd. in 1997 and until 2007 was a research engineer in the company's systems development laboratory. He is currently an advisor to the director general of the National Institute of Informatics (NII), a professor in NII's Information and Society Research Division, and a professor in the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Japan. He is also a visiting professor at the Tsuda University, Japan, and was a visiting professor at the University of Freiburg, Germany, in 2010 and at the University of Halle-Wittenberg, Germany, in 2011. He is currently engaged in research on information security and content security and privacy. He received the Best Paper Award from the IPSJ in 2005 and 2014, the Fujio Frontier Award and the Image Electronics Technology Award in 2010, the One of the Best Papers Award from the Information Security and Privacy Conference in 2011, the IPSJ Nagao Special Researcher Award in 2011, the DOCOMO Mobile Science Award in 2014, the Information Security Cultural Award in 2016, and the IEEE Workshop on Information Forensics and Security Best Paper Award in 2017. He was a member of the Information Forensics and Security Technical Committee and the IEEE Signal Processing Society. He is the Japanese representative on IFIP TC11 (Security and Privacy Protection in Information Processing Systems).

Noboru Babaguchi received the B.E., M.E. and Ph.D. degrees in communication engineering from Osaka University, in 1979, 1981 and 1984, respectively. He is currently a professor and the dean of Graduate School of Engineering, Osaka University. His research interests include image/video analysis, multimedia computing and intelligent systems. Recently, he has been engaged in privacy protection for visual information, and security and fabrication of multimedia. He has published over 250 journal and conference papers and several textbooks. He served as a Guest Editor of IEEE TIFS, Special Issue on Intelligent Video Surveillance for Public Security & Personal Privacy. He also served as a General Co-chair of MMM2008, a General Co-Chair of ACM Multimedia 2012, a Track Co-Chair of ICPR2012, an Area Chair of IEEE ICME2013, and an Honorary Co-Chair of ACM ICMR2018. He is a Fellow of IEICE, a vice president of ITE, a Senior Member of IEEE, a member of ACM, IPSJ and JSAI.

Junichi Yamagishi received a Ph.D. by Tokyo Institute of Technology in 2006. He was a senior research fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, U.K., from 2006 to 2013. He is currently a professor at National Institute of Informatics in Japan. He was awarded the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, and the Young Scientists' Prize from the Minister of Education, Science and Technology, the JSPS prize, and DOCOMO prize in 2010, 2013, 2014, 2016, and 2018, respectively. He served previously as co-organizer for the bi-annual ASVspoof special sessions at INTERSPEECH 2013-9, the bi-annual Voice conversion challenge at INTERSPEECH 2016 and Odyssey 2018, an organizing committee member for the 10th ISCA Speech Synthesis Workshop 2019 and a technical program committee member for IEEE ASRU 2019. He also served as a member of the IEEE Speech and Language Technical Committee, as an Associate Editor of the IEEE/ACM TASLP and a Lead Guest Editor for the IEEE JSTSP SI on Spoofing and Countermeasures for Automatic Speaker Verification. He is currently a PI of JST-CREST and ANR supported VoicePersona project. He also serves as a chairperson of ISCA SynSIG and as a Senior Area Editor of the IEEE/ACM TASLP.

Naoko Nitta received the B.E., M.E., and Ph.D. degrees in Engineering from Osaka University, in 1998, 2000, and 2003, respectively. She is currently an Associate Professor in Graduate School of Engineering, Osaka University. From 2002 to 2004, she was a research fellow of the Japan Society for the Promotion of Science. From 2003 to 2004, she was a Visiting Scholar at Columbia University. Her research interests are in the areas of multimedia content and social media analysis.



Yuta Nakashima received the B.E. and M.E. degrees in communication engineering and the Ph.D. degree in engineering from Osaka University, Osaka, Japan, in 2006, 2008, and 2012, respectively. From 2012 to 2016, he was an Assistant Professor at the Nara Institute of Science and Technology. He is currently an Associate Professor at the Institute for Datability Science, Osaka University. His research interests include computer vision and machine learning and their applications.

Kazuaki Nakamura received the B.S. degree in Engineering from Kyoto University in 2005, and the M.S. and Ph.D. degrees in Informatics from Kyoto University in 2007 and 2011, respectively. He is currently an Assistant Professor at Graduate School of Engineering, Osaka University, from 2012. His research interests include image processing, image recognition, and video analysis. He is a member of IEEE, IEICE, IPSJ, and ITE.

Kazuhiro Kono received the B.E, M.E., and Ph.D. degrees in communication engineering from Osaka University, Japan, in 2005, 2007, and 2010, respectively. He is currently an Associate Professor in the Faculty of Societal Safety Sciences, Kansai University. His research interests include information security and privacy or personal information technologies. He is a member of IEICE, IPSJ, REAJ, IEEE, and ACM.

Fuming Fang received B.S. degree in automation mechanic from Changchun University, Changchun, China, in 2008. He received M.S. degree in informatics from Chiba University, Chiba, Japan, in 2013, and received Ph.D. degree in informatics in Tokyo Institute of Technology, Tokyo, Japan, in 2017. After that, he joined National Institute of Informatics as a project researcher. His research interests include information security, machine learning, voice conversion, speech synthesis, speaker recognition, speech recognition, and neural language processing.

Seiko Myojin received a Ph.D. degree in engineering from Osaka University, Japan. She is currently a specially appointed assistant professor in the Graduate School of Engineering, Osaka University, Japan. Her current research interests include phenomena caused by media surrounding human. She is qualified as a senior virtual reality specialist (VRSJ). She is currently a member of the IPSJ, IEICE, VRSJ, SICE, and HIS.

Zhenzhong Kuang received M.S. and Ph.D. degrees from China University of Petroleum, Qingdao, China in 2013 and 2017, respectively. From 2015 to 2017, he was with University of North Carolina at Charlotte, Charlotte, USA. From 2018 to 2019, he was a Researcher with the Media Integrated Communication Lab., Graduate School of Engineering, Osaka University, Osaka, Japan. He currently works in with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His research interests include image privacy protection, multimedia analysis and machine learning.

Huy H. Nguyen received B.S. degree in Information Technology from VNUHCM - University of Science, Ho Chi Minh City, Vietnam in 2013. He is currently pursuing a Ph.D. degree in computer science at the Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan. Her current research interests include security and privacy in biometrics and machine learning.

Ngoc-Dung T. Tieu received B.S. degree in Information Technology from Hanoi University of Science and Technology, Hanoi, Vietnam in 2003 and M.S. in Electronics and Computer Engineering from Korea University, Seoul, Korea in 2006. During 2008-2016, she was a lecturer at University of Transport and Communications, Hanoi, Vietnam. She is currently pursuing a Ph.D. degree in computer science at the Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan. Her current research interests include information security, machine learning and image processing.