Trusted Media Challenge Dataset and User Study

Weiling Chen†
weiling@aisingapore.org
AI Singapore
Singapore

Stefan Winkler†
winkler@nus.edu.sg
AI Singapore
Singapore

Sheng Lun Benjamin Chua
benjaimchua95@hotmail.com
AI Singapore
Singapore

See-Kiong Ng
seekiong@nus.edu.sg
AI Singapore
Singapore

ABSTRACT
The emergence of fake media that can be easily created by technology has the potential to generate potent misinformation causing harm to both society and individuals. To tackle the issue, we have organized the Trusted Media Challenge (TMC) to explore how Artificial Intelligence (AI) technologies could be leveraged to combat fake media. To enable further research, we are releasing the dataset from the TMC, consists of 4,380 fake and 2,563 real videos, with various video and audio manipulation methods employed to produce different types of fake media. We have also carried out a user study to demonstrate the quality of the TMC dataset and to compare the performance of humans and AI models. The results show that the TMC dataset can fool human participants in many cases. The TMC dataset is available for research purposes upon request via tmc-dataset@aisingapore.org.

CCS CONCEPTS
• Computing methodologies → Computer vision tasks; • Applied computing → Computer forensics.

KEYWORDS
datasets; deepfake detection; computer vision

1 INTRODUCTION
Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else’s likeness [1]. They are generated by face swapping or face reenactment using deep learning techniques. These Deepfakes with high authenticity pose new threats and risks in the form of scams, fraud, disinformation, social manipulation, or celebrity porn. In particular, fake audiovisual media such as news and interviews can be used to spread misinformation that may cause social fissures.

This has inspired us to organize the Trusted Media Challenge (TMC) [2], a five-month long competition for AI enthusiasts and researchers from around the world, to design and test out AI models and solutions that can effectively detect fake media created by various state-of-the-art Deepfake technologies. In particular, we focused on the detection of audiovisual fake media, where either both video and audio modalities may be modified. For further research by the community, we are releasing the TMC dataset which includes 4,380 fake and 2,563 real videos that were part of the challenge.

High-quality fake video clips may include elaborate manipulations in both video (i.e. image frames) and audio tracks. We believe an audiovisual Deepfake dataset such as ours is useful to the research community. Another important feature of the TMC dataset is its focus on Asian content and ethnicities, whereas other datasets usually have a majority of Caucasians. Given that many Deepfake detectors are sensitive to skin tones and facial features, the TMC dataset serves as an important supplement to existing datasets and can help researchers develop more robust (and potentially less biased) Deepfake detectors.1

The remainder of the paper is organized as follows. Section 2 reviews the related work on existing Deepfake datasets and user studies. Section 3 presents details of the construction of the TMC dataset. Section 4 describes the experimental setting and results of the user study. Section 5 introduces the Trusted Media Challenge and the winning models. The last section concludes the paper.

1 While the TMC dataset has been intentionally designed for evaluating the performance of detection models on Asian demographics, using the TMC dataset on its own to develop detection models may lead to potential bias towards Asians, just as using existing datasets focused on Caucasians or other ethnicities can lead to the corresponding biases [3].
2 RELATED WORK

2.1 Deepfake Datasets

Building large datasets of high quality for fake media detection requires much effort on data collection and media manipulation. There are some existing Deepfake datasets focusing on video manipulation, including FaceForensics++ (FF++) [4], DFD [5], Celeb-DF [6], DF-1.0 [7], DFDC [8], WildDeepfake [9] and KoDF [10]. Except for DFDC, all the other datasets only include video clips without audio. Even for the DFDC dataset, although the authors claim they have done some manipulations on audios, details are not disclosed and such manipulations are not actually leveraged in the challenge. In addition, only DFD, DFDC and KoDF have explicit consent from all the actors in their videos. Moreover, the ethnicity distribution is not clear in many datasets, but it seems most of them have a majority of Caucasians and lack Asian subjects (e.g. 5.1% in celeb-DF and 12% in DFDC preview dataset [11]). One exception is KoDF [10], a deepfake dataset focusing only on Koreans.

There are far fewer existing datasets for fake audio detection. One of the most popular ones is the series of ASVspoof datasets [12–14]. Among them, only ASVspoof 2015 specifically focused on the detection of synthetic and converted speech.

A summary of the different datasets is shown in Table 1.

2.2 User Studies

Although there are numerous face forgery datasets, there are not many user studies on these datasets. FaceForensics++ conducted a user study with 204 participants (mostly computer science students) on their dataset. The authors randomly set a time limit of 2, 4, or 6 seconds and then asked the attendees whether the displayed image selected from their test set was ‘real’ or ‘fake’. Each attendee was asked to rate 60 images, which resulted in a collection of 12,240 human decisions.

Deep Video Portraits [15] and DeeperForensics [7] designed a user study on real/fake videos. 100 professionals who specialized in computer vision research were engaged. Participants were then required to rate each clip on a scale from 1 to 5 according to its “realness”. The authors assume that participants who give a score of 4 or 5 believe the video is real.

Again these user studies focused on human performance on the detection of fake videos rather than fake audio detection in the audiovisual setting. To fill this gap, we designed our user study that considers both video and audio modalities.

3 TMC DATASET

3.1 Data sources

We collected real-life footage to construct the TMC dataset from several different sources, including:

- CNA\(^2\) (Presenters and journalists talking in news programs)
- ST\(^3\) (Interviewees answering questions in TV interviews)
- Freelancers (people talking about different topics like hobbies, movies, food, etc.)

We obtained authorization to use and edit the footage produced by CNA and ST to generate and use fake clips for the purposes of the challenge. Similarly, we obtained consent from the freelancers to use and edit such clips to generate fake media. All the videos obtained from different sources are converted to mp4 format, and the resolution is changed to 360p or 1080p. More details about the collection and pre-processing of the video footage collected from the 3 sources are provided in Appendix A.1.

3.2 Generation Methods

Based on the real videos collected from the above three sources, the fake media are created using different fake video and/or audio generation methods.

3.2.1 Video Manipulation. The TMC dataset includes several fake video generation methods such as Deepfakes [4], FSGAN [16] and FaceSwap [4]. We chose these methods as they were the most popular video manipulation methods at the time the dataset was created.

3.2.2 Audio manipulation. We used StarGAN-VC [17] and One-Shot VAE [18] as audio manipulation methods to generate fake audio for the TMC dataset. A brief introduction of each video and audio generation method is provided in Appendix A.2.

3.2.3 Lip synchronization error. To create videos with a mismatch between video and audio content, we take the video footage from one real video and the audio segment from another real video and combine them into a new clip. However, we impose the restriction that both the audio and video come from persons with the same gender.

3.3 Perturbations

To simulate transmission errors, compression as well as different visual and audio effects, we add perturbations to both video and audio tracks. Detailed proportion of each perturbation method and sample frames are provided in Appendix A.3.

3.3.1 Video Perturbations. We selected 12 video perturbation methods to apply on the TMC dataset, which can be divided into 3 categories:

- Weather effects (fog, snow, etc.)
- Lighting effects (brightness change, contrast change, etc.)
- Others (compression, scaling, shaky camera, etc.)

Lighting and shaky camera effects account for a slightly higher percentage compared to other perturbation methods.

We utilize imgaug\(^4\) to add perturbations to videos in the TMC dataset. For each original real or fake video which needs to be perturbed, we randomly create 0 to 4 copies and chose one perturbation method to apply on each copy. The parameters for each perturbation are randomly set within a controlled range.

To avoid misuse of the TMC dataset, we have added a watermark to all videos. This is not regarded as a perturbation but may still have an impact on some fake detectors.

3.3.2 Audio Perturbations. We selected 3 audio perturbation methods to apply on the TMC dataset, which are volume change, additive

\(^{2}\)CNA: https://www.channelnewsasia.com/
\(^{3}\)ST: https://www.straitstimes.com/
\(^{4}\)imgaug: https://imgaug.readthedocs.io/en/latest/
Table 1: Deepfake datasets

| Dataset    | Year released | Total clips | Consented subjects | Video methods | Audio methods | Video perturbs | Audio perturbs |
|------------|---------------|-------------|--------------------|---------------|---------------|---------------|---------------|
| FF++       | 2019          | 5,000       | NA                 | 4             | 0             | 0             | 0             |
| DFD        | 2019          | 3,000       | 28                 | 5             | 0             | 0             | 0             |
| Celeb-DF   | 2019          | 6,229       | NA                 | 1             | 0             | 0             | 0             |
| DF-1.0     | 2020          | 60,000      | 100                | 1             | 0             | 7             | 0             |
| DFDC       | 2020          | 128,154     | 960                | 8             | 1             | 16            | 0             |
| WildDeepfake | 2020       | 7,314       | NA                 | NA            | 0             | 0             | 0             |
| KoDF       | 2021          | 62,166      | 403                | 6             | 0             | 0             | 0             |
| ASVspoof2015 | 2015       | 16,375      | NA                 | 0             | 10            | 0             | 0             |
| TMC        | 2022          | 6,943       | 181                | 4             | 2             | 12            | 3             |

1 Number of subjects who have given their consent to use and edit their videos.
2 Number of fake video generation methods used.
3 Number of video perturbation methods applied.

Gaussian noise, and frequency masking, in almost equal proportion. We utilize nlpaug\(^5\) to add perturbations to audios in the TMC dataset. Similar to video perturbations, parameters are initialized randomly within a controlled range.

3.4 Dataset Content

There are four types of fake media in the TMC dataset, as listed in Table 2. The last column shows the percentage of each fake type in the TMC dataset. Type-4 represents lip sync error described above — although both video and audio are real, the speech content does not match the mouth movement in the videos. Therefore, it is regarded as fake as well.

Table 2: Different types of fake media in the TMC dataset.

| Type | Video | Audio | Percentage |
|------|-------|-------|------------|
| 1    | Fake  | Fake  | 20.24%     |
| 2    | Fake  | Real  | 22.97%     |
| 3    | Real  | Fake  | 9.07%      |
| 4    | Real  | Real  | 36.92%     |

The training dataset for the challenge did not include the detailed labels of different fake types but such labels are available together with the TMC dataset upon request.

The TMC dataset includes a training set and two hidden test sets. The training set is released together with the challenge. It consists of 4,380 fake and 2,563 real videos sourced from 181 subjects in total. As mentioned, the TMC dataset differs from many other public datasets as it focuses on the Asian ethnic group, with 72.65% of subjects Asians and 45.82% of female subjects. The videos have a variety of durations, with a minimum length of 10 seconds. Additionally, both high (1080p) and low (360p) resolution videos are included in the training set.

The test set is kept hidden to evaluate the performance of the detection models submitted as part of the challenge. The challenge was conducted in two phases, with a separate test set for each phase. The ratio of real and fake videos of both test sets are similar to that of training set. However, the test set in Phase II (2,950 videos) is larger than that of Phase I (1,000 videos) and also richer in distribution compared to the training set and the test set used in Phase I. This was done to better simulate real-world scenarios for Deepfake detection.

4 USER STUDY

Our user study design was inspired by [4, 7]. We randomly selected 5,000 videos from the TMC dataset, FaceForensics++, DeeperForensics-1.0 as well as DFDC dataset, with a proportion of 70%, 10%, 10%, 10% respectively. Each video in the user study dataset was viewed by at least 3 participants.

We recruited 176 participants, most of whom were university students. Each participant viewed and rated a set of 100 videos with various distributions of real/fake videos. For each video, participants were given 6, 8 or 10 seconds to view/listen before being redirected to the rating page. For videos sampled from the TMC dataset, which have both video and audio tracks, the participants were asked to give feedback via a 5-level Likert scale (strongly agree, agree, unsure, disagree, strongly disagree) to the following statements:

- This video looks real to me.
- This audio sounds real to me.
- Overall, this clip seems real to me.

For a video-only clip sampled from the other datasets, the participants were only asked to give feedback on video.

4.1 Comparisons with Other Datasets

The results of the user study are shown in Table 3. We assume that participants who indicated "strongly agree" or "agree" believed the video/audio is real, as also shown in the table. Since other datasets only include video manipulation, we only consider Type-1 and Type-2 fake media in TMC for comparisons in this table.

It can be observed from Table 3 that fake videos in TMC look more realistic than those in FF++ and DF-1.0, but not as realistic as those in DFDC. One possible reason is that the TMC dataset has different fake video types. For example, other datasets do not contain Type-4 (i.e. inconsistency in lip movement and speech

5 nlpaug: https://github.com/makcedward/nlpaug
We have further summarized participants’ performance in detecting Area Under the Curve (AUC) to denote human performance. Participants’ ratings - Strongly agree, agree, unsure, disagree, strongly disagree were converted to probabilities of 0, 0.25, 0.5, 0.75, 1 respectively to calculate AUC.

The AUC of human detection for different ratios of real/fake videos ranges from 0.840 to 0.895 without obvious trends. This indicates that the proportion of fake videos does not affect human judgement significantly. In comparison, the best AI model submitted for the Challenge achieved an AUC of 0.9853.

4.2 Exploring Generation Methods, Fake Types and others

We have further summarized participants’ performance in detecting fake media in terms of different generation methods and fake types. Results are shown in Tables 4 and 5.

Table 3: Distribution of user ratings for fake videos in different datasets.

|         | strongly agree | agree | unsure | disagree | strongly disagree | “real” |
|---------|----------------|-------|--------|----------|-------------------|--------|
| DF-1.0  | 0.007          | 0.067 | 0.073  | 0.234    | 0.619             | 0.074  |
| DFDC    | 0.298          | 0.348 | 0.074  | 0.136    | 0.144             | 0.646  |
| FF++    | 0.005          | 0.026 | 0.042  | 0.166    | 0.760             | 0.031  |
| TMC     | 0.054          | 0.146 | 0.085  | 0.239    | 0.476             | 0.200  |

It can be observed from the tables that participants can easily spot mismatched audio and video, making Type-4 the easiest type to be detected, followed by Type-1 in which both videos and audios are manipulated – presumably, as long as humans can detect some artifacts in either audio or video track, they would label the video as fake. Type-3 is the most difficult type for participants to detect, which indicates that people are more sensitive to video manipulation than audio manipulation.

In terms of video manipulation methods, Deepfakes-512 produces the most realistic videos. For audio manipulation, StarGAN-VC performs better than One-Shot VAE. In addition, the sources of target videos also have an impact on human performance. If both videos come from the same source, it is harder for participants to detect the fake videos (AUC 0.74), while it is easier if the videos are from different sources (AUC 0.79).

Different perturbations have similar impact on human performance. However, more videos are rated as fake regardless of their labels. This indicates that videos are more likely to be rated as fake when they are distorted in some way, but it does not become easier for participants to detect fake videos with perturbations present.

Table 4: User performance in detecting different types of fake media

| Fake Type | AUC score |
|-----------|-----------|
| Type-1    | 0.914     |
| Type-2    | 0.864     |
| Type-3    | 0.769     |
| Type-4    | 0.930     |
| Overall   | 0.870     |

Table 5: User performance in detecting fake media generated by different methods

| Method            | AUC score |
|-------------------|-----------|
| Deepfakes-256     | 0.890     |
| Deepfakes-512     | 0.810     |
| Faceswap          | 0.943     |
| FSGAN             | 0.915     |
| One-Shot VAE      | 0.930     |
| StarGAN-VC        | 0.849     |

5 TRUSTED MEDIA CHALLENGE

The Trusted Media Challenge was hosted by AI Singapore from July to December 2021 with a total prize pool of up to SGD700K (approximately USD500K). In total, there were 589 participants forming 475 teams registered for the challenge. The challenge consisted of two phases with different test sets, as described in Section 3.4.

The 3 winning solutions are based on modular detection of the different types of manipulations presented in the competition including video manipulation, audio manipulation as well as inconsistency in lip movement and speech content. The final result was calculated by combining individual detection scores from different models using specific aggregation strategies. The best team achieved an AUC of 0.9853. Further details of the top 3 models and their performance are provided in the Appendix B.

6 CONCLUSIONS

In this work, we presented a fake media dataset consisting of 6,943 fake and real videos generated with a variety of manipulation methods. Different perturbations were added to both real and fake videos to increase the difficulty of detection. The dataset was used in the Trusted Media Challenge held in 2021 by AI Singapore. We have designed and carried out a comprehensive user study to demonstrate the quality of TMC dataset and explored the impact on human performance with regards to different factors. Our study results indicated that the TMC dataset can fool humans in many cases. However, based on the results from the Trusted Media Challenge, we also found that AI models can beat humans for Deepfake detection on our TMC dataset. Due to limited space, the full appendix is available online at https://arxiv.org/abs/2201.04788

ACKNOWLEDGMENTS

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme.
REFERENCES

[1] Wikipedia. Deepfakes. https://en.wikipedia.org/wiki/Deepfake, 2022. Accessed Apr 30, 2022.

[2] Abigail Toh. Aisg launches "prize challenge" to curate ideas and ai models to combat fake media. https://asingapore.org/2021/07/aisg-launches-prize-challenge-to-curate-ideas-and-ai-models-to-combat-fake-media/, 2022. Accessed March 7, 2022.

[3] Simone Fabbriuzzi, Syrene Papadopoulos, Eirini Nioutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. arXiv preprint arXiv:2107.07919, 2021.

[4] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proc. IEEE/CFV International Conference on Computer Vision (ICCV), pages 1–11, 2019.

[5] Nick Dufour and Andrew Oudl. Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html, 2022. Accessed Apr 30, 2022.

[6] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In Proc. IEEE/CFV Conference on Computer Vision and Pattern Recognition (CVPR), pages 3207–3216, 2020.

[7] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In Proc. IEEE/CFV Conference on Computer Vision and Pattern Recognition (CVPR), pages 2889–2898, 2020.

[8] Brian Dollhansky, Joanna Bitton, Ben Pillam, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake detection challenge (DFDC) dataset. arXiv preprint arXiv:2006.07597, 2020.

[9] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wild-D: A large-scale dataset for deepfake images. In Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pages 10744–10753, 2021.

[10] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Wang, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proc. IEEE/CFV International Conference on Computer Vision (ICCV), pages 7184–7193, 2019.

[11] Hikokura Kameoka, Takahiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In Proc. IEEE Spoken Language Technology Workshop (SLT), pages 266–273. IEEE, 2018.

[12] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. arXiv preprint arXiv:1904.05742, 2019.

[13] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Perez, Christian Richardt, Michael Zollhofer, and Christian Theobalt. Deep video portraits. ACM Transactions on Graphics (TOG), 37(4):1–14, 2018.

[14] Mamunulhossain Siddiqui, Md Sahidul Islam, Héctor Delgado, Andreas Nauck, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In Proc. Interspeech, 2019.

[15] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pages 7184–7193, 2019.

[16] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nauck, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In Proc. Interspeech, 2019.

[17] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Perez, Christian Richardt, Michael Zollhofer, and Christian Theobalt. Deep video portraits. ACM Transactions on Graphics (TOG), 37(4):1–14, 2018.

[18] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pages 7184–7193, 2019.

[19] Hikokura Kameoka, Takahiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In Proc. IEEE Spoken Language Technology Workshop (SLT), pages 266–273. IEEE, 2018.