ILLUSION: UNVEILING TRUTH WITH A COMPREHENSIVE MULTI-MODAL, MULTI-LINGUAL DEEPFAKE DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

The proliferation of deepfakes and AI-generated content has led to a significant increase in media forgeries and misinformation, necessitating development of more robust detection systems. Current datasets, however, lack comprehensive diversity across modalities, languages, and real-world scenarios. To address this gap, we present ILLUSION (Integration of Life-Like Unique Synthetic Identities and Objects from Neural Networks), a large-scale multi-modal deepfake dataset comprising over 1.3 million samples. ILLUSION encompasses (i) audio-visual forgeries, (ii) diverse linguistic content with over 26 languages, (iii) challenging noisy environments, and (iv) various manipulation protocols. Generated using state-of-the-art generative models, ILLUSION includes face swaps, audio spoofing, synchronized audio-video manipulations, and synthetic images, faces, and videos. The proposed dataset has balanced representation of gender and skin tone, supports multilingual experiments, and is designed to facilitate development of robust multi-modal detection systems. We benchmarked state-of-the-art algorithms across multiple modalities including image-based, audio-based, video-based, and multi-modal detection. The results highlight critical challenges such as (a) performance degradation in multi-lingual and multi-modal contexts, (b) accuracy reduction in noisy environments, and (c) limited generalization to real-world scenarios and zero-day attacks. It is our assertion that the comprehensive nature of the proposed dataset enables researchers to develop and evaluate more resilient deepfake detection methods, addressing the evolving landscape of synthetic media threats.

1 INTRODUCTION

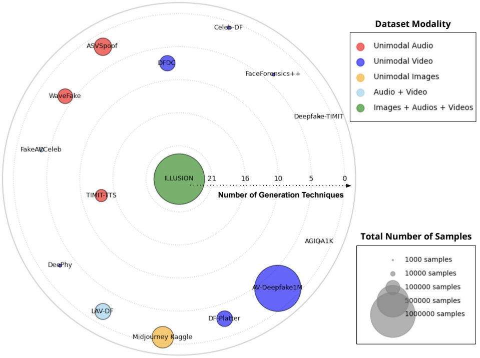

Figure 1: Comparative analysis of the proposed dataset with existing ones based on modalities, size, and manipulations.

The emergence of social media platforms has fundamentally transformed our mode of communication and information dissemination. Platforms such as Facebook, YouTube, Instagram, and TikTok, which boast billions of users worldwide, have expanded the scope of shared content beyond text to include images, videos, and other forms of multimedia. This shift has precipitated a surge in the volume of multi-modal content accessible online. As social networks evolve rapidly, they have emerged as the primary conduit for disseminating user-generated multi-modal content. The data circulating on these social networks is predominantly multi-modal, encompassing videos, audio, and images. With their billions-strong user base, these platforms generate enormous data every minute. Nonetheless, the rise of social media and multi-modal content has concurrently fueled an upsurge in the spread of deepfakes and synthetic media fabricated by deep learning techniques. The advancements in generative techniques like Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAEs), and diffusion-based models, have significantly enhanced the realism of synthetically generated content, making it more convincing to the untrained eye. These AI foundational models and diffusion-based Generative AI (GenAI) models have exhibited unparalleled competence in comprehending and generating human-like videos,

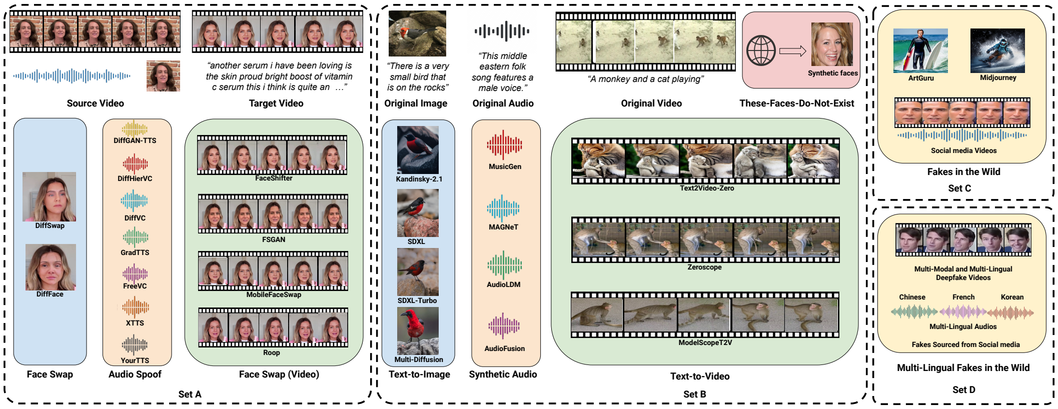

Figure 2: Visual representation and organization of each subset in the proposed ILLUSION dataset.

images, and sounds. In 2023, deepfake videos online reached 95,820, which is a 550% increase from 2019[1]. Forecasts suggest that by 2026, up to 90% of online content could be synthetically generated[2]. The accessibility of these models raises concerns about this prevalent misuse. Therefore, developing deepfake detection techniques is crucial. Many researchers propose various detection strategies for identity-based deepfakes (Afchar et al., 2018; Nguyen et al., 2019; Agarwal et al., 2019; Khalid & Woo, 2020; Chhabra et al., 2023). High-quality deepfake datasets are essential for effective detection methods. However, existing datasets focus mainly on identity-swap deepfakes and overlook multi-modal, multi-lingual and AIGC content. There is a lack of datasets with entirely fabricated data across images, audio, and video. This gap hinders progress in multi-modal and multi-lingual deepfake detection research.

To bridge this dataset gap, we introduce *ILLUSION*: **I**ntegration of **L**ife-**L**ike **U**nique **S**ynthetic **I**dentities and **O**bjects from **N**eural Networks[3], a novel multi-modal, multi-lingual deepfake dataset divided across four sets (visualized in Figure 2). Set A is an identity forgery dataset with audio-video synchronized. Set B incorporates AI-generated synthetic data covering three media modalities: image, audio, and video. Set C, a test set, includes a pool of real-world AI-generated content (AIGC) sampled from different sources and set D includes multi-lingual and multi-modal deepfake samples spanning over 26 different languages. The dataset is prepared with continuous usage of 40 GPUs, accounting for 2000 GB of cumulative memory. With over 800 GBs in size, the dataset contains over 1.3 million samples encompassing the four sets. To the best of our knowledge, this is one of the largest datasets containing vast variability of generation methods, different modalities, multiple languages, and various challenges (refer Figure 1).

The proposed comprehensive dataset provides diverse AI-generated content to serve as a valuable asset for research in detecting AI-generated media varying in input modality, generation models, different languages, and content type. To assess and analyze the utility of our dataset, we conduct extensive experiments and benchmark using 11 baseline deepfake methods and analyze their performance when tested in different settings. The primary contributions of our work are summarized below:

- We introduce a multi-modal deepfake dataset developed using 28 GenAI models grounded in GANs, VAEs, Transformers, and Diffusion-based models, spanning image, audio, and video modalities. This dataset is partitioned into four distinct sets.
- The dataset encompasses identity manipulations, where forgery can manifest across audio, video, or both. This set is seamlessly synchronized across audio-visual channels and maintains a balance in terms of sex and skin tone.
- The dataset also includes AI-generated content (AIGC) produced by various text-to-modality models across image, audio, and video domains. It encompasses a subset of entirely synthetic faces. Additionally, the dataset features real-world deepfakes, designed to evaluate detection algorithms in a practical context that spans multiple modalities and languages.

---

[1] Deepfake Statistics – Current Trends, Growth, and Popularity

[2] How can we combat the worrying rise in the use of deepfakes in cybercrime

[3] Dataset Webpage: `https://anonymousillusion.github.io/glowing-sniffle/`

Table 1: Details of publicly available deepfake datasets.

Dataset Name	Year	Real Samples	Fake Samples			Total Samples	Generation Techniques	Identity Swapping	AI Generated Content	Multi-Lingual	AI-Swap-Lingual
			Images	Audio	Video						
Deepfake-TIMIT (Korshunov, 2018)	2018	640	N/A	N/A	320	960	2	✓	×	×	×
FaceForensics++ (Rossler et al., 2019)	2019	1000	N/A	N/A	4,000	5000	4	✓	×	×	×
Celeb-DF (Li et al., 2020)	2020	590	N/A	N/A	5,639	6,229	1	✓	×	×	×
DFDC (Dolhansky et al., 2020)	2020	23,654	N/A	N/A	104,500	128,154	8	✓	×	×	×
DeeperForensics-1.0 (Jiang et al., 2020)	2020	50,000	N/A	N/A	10,000	60,000	1	✓	×	×	×
ASVSpoof (Yamagishi, 2019)	2021	16,492	N/A	148,148	N/A	164,640	19	✓	×	×	×
WaveFake (Joel Frank, 2021)	2021	0	N/A	117,985	N/A	117,985	6	✓	×	×	×
FakeAVCeleb (Khalid et al., 2021)	2021	500	N/A	500	9000	10,000	4	✓	×	×	×
TIMIT-TTS (Salvi et al., 2022)	2022	430	N/A	80,000	N/A	80,430	14	✓	×	×	×
DeePhy (Narayan et al., 2022)	2022	100	N/A	N/A	5,040	5,140	3	✓	×	×	×
LAV-DF (Cai et al., 2022b)	2022	36,431	N/A	33,176	65,997	136,304	2	✓	×	×	×
Midjourney Kaggle (Iulia Turc, 2022)	2022	0	250,000	N/A	N/A	250,000	1	×	✓	×	×
DF-Platter (Narayan et al., 2023)	2023	764	N/A	N/A	132,496	133,260	3	✓	×	×	×
AV-Deepfake1M (Cai et al., 2023)	2023	286,721	N/A	N/A	860,039	1,146,760	3	✓	×	×	×
AGIQA1K (Zhang et al., 2023)	2023	0	1,080	N/A	N/A	1,080	2	×	✓	×	×
TWIGMA (Chen & Zou, 2024)	2024	0	800,000	N/A	N/A	800,000	N/A	×	✓	×	×
ILLUSION (Proposed)	**2024**	**139,740**	**905,548**	**27,244**	**299,454**	**1,371,986**	**28**	✓	✓	✓	✓

- We benchmark the proposed dataset using state-of-the-art detection methods and conduct a comprehensive analysis of their performance across a range of challenging protocols.

2 RELATED WORKS

Deepfake detection performance is highly dependent on the quality of the dataset used in terms of the variation in the modality of deepfakes, generation techniques, and the realisticity of each sample. Also, a balanced dataset is crucial for unbiased learning and detection. Early datasets like DF-TIMIT (Korshunov, 2018), FaceForensics++ (Rössler et al., 2018), Celeb-DF (Li et al., 2020), WildDeepfake (Zi et al., 2020) and DeeperForensics-1.0 (Jiang et al., 2020) were small to medium in size and unimodal. Larger datasets like $\mathrm{FFIW}_{\mathbf{10k}}$ (Zhou et al., 2021), KoDF (Kwon et al., 2021), and DF-Platter (Narayan et al., 2023) focused on visual manipulation. Unimodal audio-based manipulation was introduced in ASVSpoof (Wang et al., 2020b), WaveFake (Joel Frank, 2021), TIMIT-TTS (Salvi et al., 2022), and multi-modal manipulation in DFDC (Dolhansky et al., 2020) and FakeAVCeleb (Khalid et al., 2021). LAV-DF (Cai et al., 2022a) and AV-Deepfake1M (Cai et al., 2023) were the first large-scale datasets with multi-modal AV manipulations but lacked non-identity-based AIGC. Most previous datasets focused on a few specific generation techniques. Our study presents a dataset overcoming these limitations by including a diverse set of generative models. It's balanced across sex and skin tone for identity forgery-based deepfakes, and includes non-identity-based synthetic samples across all domains with multiple modalities, making it one of the only datasets with partial and fully synthetic samples.

3 THE ILLUSION DATASET

In this paper, we present *ILLUSION:* **I**ntegration of **L**ife **L**ike **U**nique **S**ynthetic **I**dentities and **O**bjects from **N**eural networks, a comprehensive large-scale multi-modal deepfake dataset[4]. This dataset comprises 1,376,371 samples, spanning image, audio, video, and synchronized audio-video modalities. It stands as the largest multi-modal dataset in the current deepfake literature. The dataset is divided into four subsets: Set A, Set B, Set C, and Set D. Set A includes identity manipulations featuring faceswaps, voice spoofs, and both. Set B comprises synthetically generated media, including images and videos of sceneries, objects, situations, and music audio. This set also incorporates synthetic faces generated from the website[5]. Set C encompasses real-world testing samples, i.e., **Fakes in the Wild**, generated using proprietary generative models, and Set D is a multi-lingual multi-modal deepfake testing set. The dataset is produced using 28 distinct generative models, encompassing open-source and proprietary models. Although most publicly available deepfake datasets exhibit an imbalance in terms of sex and skin tone (Nadimpalli & Rattani, 2022; Xu et al., 2022a), the ILLUSION dataset ensures balance across both subgroups.

3.1 DATASET STATISTICS AND ORGANIZATION

This section discusses the statistics and organization of each set of the proposed ILLUSION dataset. Table 2 presents the set-wise statistics.

---

[4] The collection and generation of the ILLUSION dataset is approved by the Institutional Ethics Review Committee. The dataset will be provided only to academic institutions for research purposes

[5] This Person Does Not Exist: `https://thispersondoesnotexist.com`

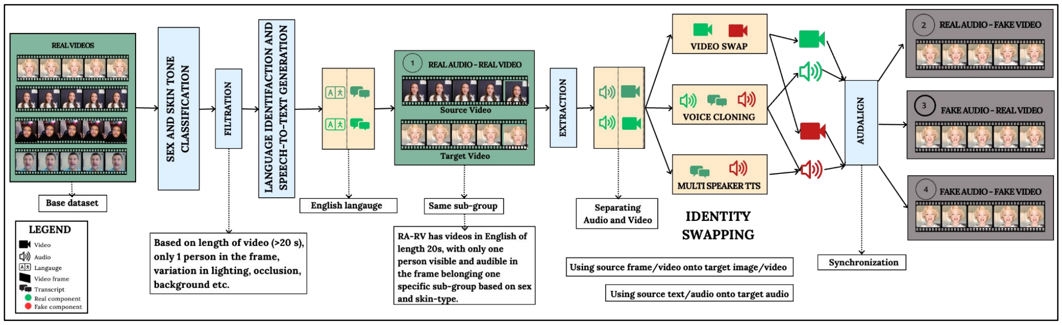
REAL VIDEOS
Base dataset
LEGEND
Video
Audio
Langauge
Video frame
Transcript
Real component
Fake component
SEX AND SKIN TONE CLASSIFICATION
FILTRATION
Based on length of video (>20 s), only 1 person in the frame, variation in lighting, occlusion, background etc.
LANGUAGE IDENTIFACTION AND SPEECH-TO-TEXT GENERATION
English langauge
1 REAL AUDIO - REAL VIDEO
Source Video
Target Video
Same sub-group
RA-RV has videos in English of length 20s, with only one person visible and audible in the frame belonging one specific sub-group based on sex and skin-type.
EXTRACTION
Separating Audio and Video
VIDEO SWAP
VOICE CLONING
MULTI SPEAKER TTS
IDENTITY SWAPPING
Using source frame/video onto target image/video
Using source text/audio onto target audio
AUDALIGN
Synchronization
2 REAL AUDIO - FAKE VIDEO
3 FAKE AUDIO - REAL VIDEO
4 FAKE AUDIO - FAKE VIDEO

Figure 3: Illustration of generation pipeline employed in set A for identity swaps.

Set A: This set comprises a total of 13 identity manipulations, generated from 200 unique identities sampled from the CelebV-Text dataset (Yu et al., 2023). Each audio, video, and audio-synchronized clip in this set is 20 seconds long. The samples in this set incorporate variations such as occlusions (e.g., hats, glasses, beards, etc.), body movements, and lighting conditions, thereby ensuring a diversity of variations in the dataset. This set is categorized into four classes: (i) Real Audio- Real Video (RA-RV), (ii) Real Audio and Fake Video (RA-FV), (iii) Fake Audio and Real Video (FA-RV), and (iv) Fake Audio and Fake Video (FA-FV) (as visualized in Figure 3). Table 2 summarizes the number of samples available in each class of Set A. Manipulations in images and videos for each of these classes are introduced by swapping the faces of the source identity onto the target video. For this purpose, we employ 6 different face-swapping models, namely, MobileFaceSwap (Xu et al., 2022b), FSGAN (Nirkin et al., 2019), FaceShifters (Li et al., 2019b), ROOP (s0md3v, 2023), DiffFace (Kim et al., 2022), and DiffSwap (Zhao et al., 2023). The audio deepfakes are created using the English transcription generated through the MMS model (Pratap et al., 2023). This transcription is then employed by Text-to-Speech systems to create identity-swapped voice clones. We utilize 7 different audio-generative models, namely, FreeVC (Li et al., 2023), XTTS (Eren & The Coqui TTS Team, 2021), DiffVC (Popov et al., 2022), DiffHierVC (Choi et al., 2023), YourTTS (Casanova et al., 2022), DiffGAN-TTS (Liu et al., 2022), and GradTTS (Popov et al., 2021) for voice-swapping. For classes RA-FV, FA-RV, and FA-FV, we employ Audalign[6], a fingerprinting-based model to ensure seamless synchronization between the audio and video, thereby enhancing the realism of the generated fake. The details of all the models utilized in this set are available in the Appendix.

Table 2: Dataset statistics of the ILLUSION dataset. In Set A, the audio and images are extracted from the original video for manipulation. Further, the videos in Set A comprises of 3 compressions (raw, C23, and C40).

Sets	**Modality**	**Generation Methods**	**Real Samples**	**Fake Samples**	**Total**
Set A	Audio	8	N/A	6,400	6,400
	Images	2	N/A	403,200	403,200
	Videos	4	1,032	278,400	279,432
Set B	Audio	4	5,211	20,844	26,055
	Images	4	118,287	473,148	591,435
	Videos	3	7,010	21,030	28,040
	Faces	1	8,200	8,200	16,400
Set C	Images	2	N/A	21,000	21,000
	Videos	N/A	N/A	24	24
Set D	Audios	N/A	1600	2560	4160
	Videos	N/A	100	125	225
	Total	**28**	**141,440**	**1,234,931**	**1,376,371**

Set B: This set comprises a total of 523,222 entirely synthetic samples and their 138,708 real counterparts, generated through 11 open-source models and one closed-source model. This set includes images, audio, and videos primarily generated using diffusion models and transformers. For the generation of synthetic images, we employ the images and their corresponding prompts from the training set of the COCO dataset (Lin et al., 2014) to generate using four text-to-image generative models. These models include Stable Diffusion-XL (Podell et al., 2023), Kandinsky 2.1 (Razzhigaev et al., 2023), MultiDiffusion (Bar-Tal et al., 2023), and SDXL-Turbo (Sauer et al., 2023). We also collected 8,200 synthetic face images from “This Person Does Not Exist", ensuring a balance in terms of sex and skin tone. These identities are entirely synthetic, have not been swapped, and do not exist in the real world. To generate synthetic audio, we utilize audios and corresponding captions from the MusicCaps dataset (Agostinelli et al., 2023) and generate 5,211 synthetic audio samples each from three text-to-audio generative models and one audio-to-audio model, namely, AudioLDM

---

[6] Audalign: `https://github.com/benfmiller/audalign`

(Liu et al., 2023), MusicGen (Copet et al., 2024), MAGNeT (Ziv et al., 2024), and Audio Diffusion [7]. Similarly, we also generated synthetic videos using three text-to-video generative models, namely, Text2Video-Zero (Khachatryan et al., 2023), ModelScopeT2V (Wang et al., 2023a), and ZeroScope [8]. For this, we borrow the corresponding caption for each video from the training set of MSRVTT dataset (Xu et al., 2016) and utilize it to generate 7,010 synthetic videos. The details of all the models utilized in this set are available in the Appendix.

Set C: This set serves as a real-world testing set, comprising 21,024 fake samples. It is a curated collection of viral deepfake videos circulated on social media platforms and samples generated using proprietary models such as MidJourney[9] and ArtGuru[10]. The former includes 24 identity-swapped videos. The latter consists of synthetic images generated through a premium API and a web interface, respectively. For Midjourney, we utilize prompts from the validation set of the COCO dataset, and for each prompt, we obtain four corresponding images, resulting in a total of 20,000 images. ArtGuru, specializes in generating identity-driven images for a given prompt. Therefore, we randomly select 1000 prompts from class “person" from COCO dataset to generate a total of 1000 images.

Set D: This set is a multi-lingual, multi-modal real-world testing set comprising 4385 samples. Curated from social media platforms, it spans over 26 different languages, including French, German, Italian, Chinese, Korean, Arabic, Japanese, Tamil, Kannada, Oriya, Hindi, Sanskrit, Latin, Punjabi, and Gujarati. Set D is divided into two parts: D.1 consists of 4160 web-curated multi-lingual samples, while D.2 is a subset of 225 multi-modal multi-lingual deepfake samples, annotated with four classes (RA-RV, RA-FV, FA-RV, and FA-FV). Additionally, the audio and video in the samples are synchronized.

3.2 SIZE AND FORMAT

The ILLUSION dataset is approximately 800 GB in size. Each clip in Set A has a duration of 20 seconds. All face images in the dataset were either synthetically generated or obtained from publicly available datasets with proper licensing and consent where applicable. The videos are provided in the MPEG4.0 format, with a resolution of 512 × 512 and the original frame rate of videos. The dataset maintains consistency across resolution, compression, and the generation technique utilized. For compression at levels c23 and c40, we employ the H.264 video compression standard. We categorize skin tones into four bins based on the Fitzpatrick scale (details available in the Appendix) and consider two sexes, resulting in eight sub-groups. To ensure high-quality swaps, identities are swapped only within the same sub-groups. In Set B, we generate 24 frames for every video from Text2Video-Zero and ZeroScope and 40 frames for each video from ModelScopeT2V. The videos in Set C, being curated from various sources, exhibit variability in resolution and length. However, all the images generated using MidJourney and ArtGuru maintain a consistent resolution of 1024 × 1024 and 512 × 512, respectively.

3.3 AUDIO AND VISUAL QUALITY ASSESMENT

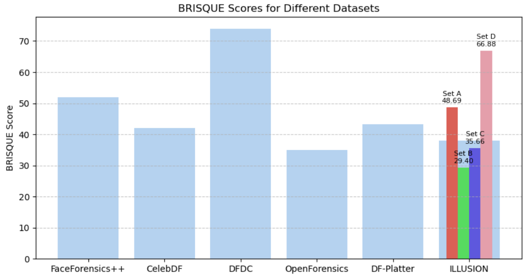

Figure 4: Comparing Brisque Score of ILLUSION with other datasets.

To evaluate the visual of the proposed dataset, we use the BRISQUE score (Mittal et al., 2012) quality metric, respectively, for all four sets as shown in Figure 4. We observe a mean Brisque Score of 38.04. On a scale of 0 (best) - 100 (worst), the average BRISQUE scores for the entire dataset and individual sets are shown in table 3. The BRISQUE scores for Face-Forensics++, CelebDF, DFDC, OpenForensics, and DF-Platter are approximated from (Narayan et al., 2023). Further, the table also includes FAD scores (Kilgour et al., 2018), quantitatively reflecting the quality of audio samples individually in both the sets and the whole dataset. We report a mean FAD of 9.43 for the proposed ILLUSION dataset. These scores highlight that the proposed dataset is of high quality and is challenging with multiple covariates.

---

[7] Audio Diffusion: `https://huggingface.co/teticio/audio-diffusion-256`

[8] ZeroScope: `https://huggingface.co/cerspense/zeroscope_v2_576w`

[9] MidJourney: `https://www.midjourney.com/home`

[10] ArtGuru: `https://www.artguru.ai/`

3.4 COMPUTATIONAL SETUP

In Set A, we utilize a total of 13 generation methods to produce identity-swaps across image, audio, video, and audio-video synchronized modalities. This process is facilitated by Nvidia A100 with 16 GPUs, each with 80GBs of memory. Set B is generated through 11 open-source and one closed-source generative models, utilizing two Nvidia A40 GPUs, each with 48GBs of memory, and three Nvidia DGX stations, each equipped with four V100 GPUs of 32GB memory. Set C comprises samples generated from two proprietary models, produced on 2 Nvidia 3090 GPUs, each with 24 GBs of memory. The benchmarking experiments for the dataset are conducted on 2 A40, each with 48GBs of memory, and 6 A30 GPUs, each with 24GBs of memory, in a multi-GPU setup.

Table 3: Audio and visual quality assessment of ILLUSION dataset.

Dataset Split	**Vision (Mean Brisque Score)**			**Audio (Mean Fréchet Audio Distance)**		
	Train Set	**Test Set**	**Overall**	**Train Set**	**Test Set**	**Overall**
Set A	49.08	42.98	48.69	6.37	7.30	6.41
Set B	29.43	29.29	29.40	6.64	6.58	6.55
Set C	N/A	35.66	35.66	N/A	N/A	N/A
Set D.1	N/A	N/A	N/A	N/A	40.39	40.39
Set D.2	N/A	66.88	66.88	N/A	N/A	N/A
	Overall: 38.04			**Overall: 9.43**		

4 EXPERIMENTAL SETUP

This section outlines the training and testing protocol established for the proposed ILLUSION dataset, followed by a discussion on the deepfake detection methods and evaluation metrics employed for benchmarking. The proposed dataset is designed to address the following pivotal research questions:

RQ1: How effective are the detection systems in detecting multi-modal identity-swaps?

RQ2: How effective are the detection systems in identifying synthetically generated media?

RQ3: How robust and reliable are the current state-of-the-art detection algorithms when deployed in real-world scenarios?

RQ4: Is it feasible to detect identity swaps and synthetic media in a zero-day attack setting?

RQ5: Is it possible to successfully trace back the source of a given deepfake?

4.1 EVALUATION PROTOCOLS

The ILLUSION dataset is composed of four sets. Sets A and B are partitioned into training and testing subsets in a ratio of 3:1. The training data is split into a 9:1 ratio to divide into train and validation data. To mitigate the skew between the “Real" and “Fake" classes in set A, we borrow an additional 144 videos (18 subjects/sub-group) from the CelebV-Text dataset. In contrast, set C and D is exclusively a test set. For all the videos in Set A, we extract 10 frames from each fake video and all from each real video. For Set B, we pick 24 frames each from the generative models and select every sixth frame from real videos. Further, for synthetic images generated from four text-to-image models, we repeat their corresponding real images four times. This approach addresses the imbalance between the dataset’s real and fake samples.

Protocol 1 - Multi-modal Deepfake Detection: This protocol utilizes Set A and Set B, each with their respective training and test sets. Set A is divided in a subject-disjoint manner, incorporating 160 subjects in the training set (20 subjects/sub-group) and 40 subjects in the test set (5 subjects/sub-group). The state-of-the-art audio, video, and multi-modal deepfake detection models are then trained and tested on the samples from Set A. The results are presented in three compression settings - raw, C23, and C40 - to facilitate the assessment of deepfake quality in the dataset relative to existing datasets (results in the Appendix). For Set B, images borrowed from the COCO dataset (Lin et al., 2014) and BFW dataset (Robinson, 2022) serve as real samples corresponding to the fake samples generated using text-to-image models and synthetic faces, respectively. Similarly, audio and video samples from the MusicCaps dataset (Agostinelli et al., 2023) and MSRVTT dataset (Xu et al., 2016) are used as real samples corresponding to the fake samples generated using text-to-audio and text-to-video models. We extract 24 frames from fake videos to classify synthetic videos and select every 6th real video frame to maintain data balance.

Protocol 2 - Zero-shot/Zero-day Generalization: The primary aim of this protocol is to test the generalizability of detections on new or unseen generation methods. The detection models are initially trained on the train set of Set A and subsequently tested on the test set of Set B. The performance of the models is also evaluated in a vice-versa setting.

Table 4: Classification performance for visual components of the dataset obtained by varying the training and testing sets.

Trained On	Models	Set A				Set B				Set C (All Fake)
		Accuracy	Accuracy-Fake	Accuracy-Real	AUC	Accuracy	Accuracy-Fake	Accuracy-Real	AUC	Accuracy
Set A	F3Net	0.851	0.751	0.951	0.945	0.468	0.475	0.460	0.462	0.341
	DSP-FWA	0.984	0.982	0.986	0.997	0.428	0.437	0.418	0.396	0.214
	MesoInceptionNet	0.505	0.999	0.882	0.883	0.505	0.988	0.473	0.487	0.991
	Xception	0.851	0.846	0.856	0.920	0.497	0.749	0.283	0.515	0.633
Set B	F3Net	0.498	0.008	0.993	0.497	0.981	0.991	0.970	0.998	0.717
	DSP-FWA	0.498	0.001	0.999	0.501	0.993	0.995	0.990	0.999	0.727
	MesoInceptionNet	0.495	0.009	0.987	0.386	0.757	0.553	0.967	0.919	0.045
	Xception	0.502	0.018	0.992	0.526	0.977	0.987	0.966	0.997	0.725
Set A + Set B	F3Net	0.881	0.926	0.836	0.958	0.956	0.986	0.925	0.994	0.703
	DSP-FWA	0.972	0.975	0.970	0.995	0.992	0.995	0.989	1.000	0.863
	MesoInceptionNet	0.481	0.013	0.425	0.701	0.834	0.806	0.912	0.948	0.241
	Xception	0.881	0.926	0.835	0.954	0.972	0.982	0.963	0.997	0.650

Protocol 3 - Generalization on Real-World Deepfake Media: This protocol assesses the performance of existing state-of-the-art models on real-world deepfake samples. Here, the models are trained on the train set of either Set A, Set B, or both, and their performance is evaluated on Set C.

Protocol 4 - Performance on Model Attribution: The final protocol presents a challenging model attribution task, i.e. to predict the generative technique used to create the input deepfake. The models are trained and tested on fake samples generated from each technique from Set A and Set B. The detection models are evaluated separately for image, video, and audio modalities.

4.2 BENCHMARKING DETAILS

DeepFake Detection Methods We utilize four state-of-the-art video and four audio deepfake detection models to benchmark all three sets of the proposed dataset. For video deepfake detection, we employ MesoInceptionNet (Afchar et al., 2018), XceptionNet (Chollet, 2017), DSP-FWA (Li & Lyu, 2019), and F3Net (Wei et al., 2020). For audio deepfake detection, we use RawGAT-ST (Tak et al., 2021), AASIST (Jung et al., 2022), SSLModel (Tak et al., 2022), and Conformer (Gulati et al., 2020). We also benchmark the proposed dataset using multi-modal deepfake detection algorithms. Specifically, we employ state-of-the-art methods such as MRDF (Zou et al., 2024) and FACTOR (Reiss et al., 2023) from the literature. Additionally, we use an ensemble of F3Net and SSLModel, which are baseline unimodal models (referred to as unimodal ensembling), and report class-wise video-level accuracy. Benchmarking the proposed ILLUSION dataset with 11 baseline algorithms provides a comprehensive evaluation, encompassing both typical methods and type-complete approaches. The unimodal baselines focus on modality-specific behaviors, enabling a deeper understanding of how state-of-the-art algorithms perform within their respective domains (e.g., image, audio, or video). In contrast, the multimodal baselines evaluate type-complete methods, capturing the interplay between multiple modalities and offering insights into cross-modal generalization and robustness. This dual benchmarking strategy ensures a balanced assessment of both specialized and holistic detection capabilities. Detailed descriptions of all these algorithms are provided in the the Appendix.

Evaluation Metrics For models trained on image and video data, we provide frame-level accuracy and Area Under the Receiver Operating Characteristic (AUC) scores. Each frame in a video is computed and classified as either fake or real. We also present class-wise accuracy for additional analysis. For audio data, we report the Equal Error Rate (EER) and AUC score. For models trained on multi-modal data, such as combined video and audio, we provide video-level accuracy, using a threshold set at 50% of frames to classify a video as fake.

Implementation Details This section provides the details of the implementation of the benchmarking experiments to ensure reproducibility. The DSFD detector (Li et al., 2019a) is used to extract faces from the frames of videos containing faces. For all protocols, the models are trained for 30 epochs with early stopping, and the models with the best validation accuracy are selected. We use the Adam optimizer with an initial learning rate of 0.0001. A batch size of 256 is used for distributed training.

5 RESULTS AND DISCUSSION

This section discusses the benchmark results obtained using the state-of-the-art deepfake detection models mentioned in Section 4.2 when trained and evaluated on the proposed ILLUSION dataset. The performance analyzed is in accordance with the protocols described in section 4.1.

Protocol 1 - Multi-Modal Deepfake Detection: To analyze the performance of audio and visual detection models, we trained and tested them on both Set A and Set B of the proposed ILLUSION

Table 5: Classification performance for audio components of the dataset obtained by varying the training and testing sets.

Trained On	**Models**	**Set A**				**Set B**			
		EER	**Accuracy-Fake**	**Accuracy-Real**	**AUC**	**EER**	**Accuracy-Fake**	**Accuracy-Real**	**AUC**
Set A	RawGAT-ST	0.443	1.000	0.000	0.576	0.470	0.999	0.004	0.539
	AASIST	0.049	0.879	1.000	0.991	0.430	0.240	0.922	0.563
	Conformer	0.025	0.996	0.975	0.991	0.562	0.715	0.081	0.395
	SSLModel	0.006	0.980	1.000	1.000	0.583	0.676	0.104	0.356
Set B	RawGAT-ST	0.381	0.125	1.000	0.699	0.032	0.988	0.939	0.995
	AASIST	0.261	0.071	1.000	0.805	0.020	0.988	0.970	0.997
	Conformer	0.821	0.008	0.825	0.119	0.005	0.992	0.994	0.999
	SSLModel	0.694	0.005	0.925	0.252	0.006	0.993	0.993	0.999
Set A + Set B	RawGAT-ST	0.025	0.943	0.975	0.994	0.035	0.973	0.958	0.995
	AASIST	0.050	0.946	0.950	0.992	0.027	0.983	0.951	0.996
	Conformer	0.030	0.959	0.975	0.995	0.020	0.990	0.959	0.997
	SSLModel	0.069	0.938	0.925	0.988	0.022	0.992	0.952	0.998

dataset. From Tables 4 and 5, for set A, we observe that all architectures perform well for visual as well as audio detection models, with DSP-FWA achieving the best performance for visual data (99.3% accuracy on Set B) and SSLModel excelling in audio data with an EER of 0.006. A similar trend is visible in set B, where we observe that all the detection models, when trained on synthetic data, are able to achieve a promising detection performance.

However, performance on Set C, which includes curated real-world and compressed deepfakes, reveals significant variability. DSP-FWA achieves 21.4% accuracy when trained on Set A and tested on Set C, compared to 72.7% accuracy when trained on Set B, highlighting the challenges of generalizing from synthetic data to real-world scenarios. MesoInceptionNet shows unusually high accuracy on Set C due to its tendency to classify most inputs as fake, leading to inflated performance on the all-fake Set C.

Set C was intentionally designed as a challenging test set to mimic real-world deepfakes crafted for mass misinformation. Its inclusion underscores the need for detection models to handle diverse generative techniques and real-world complexities, highlighting the limitations of current approaches and the forward-looking design of ILLUSION in advancing deepfake detection research.

We evaluated the effectiveness of multi-modal deepfake detection methods on Set A of the proposed ILLUSION dataset. These methods were trained on audio-video synchronized samples from Set A. The performance achieved are detailed in Table 6. Our findings show that MRDF outperforms FACTOR across all classes, notably achieving an 87.1% class-wise accuracy for the FakeAudio-FakeVideo category. Conversely, FACTOR consistently underperforms, with a notable low of 15.7% accuracy for the RealAudio-RealVideo class. Unimodal Ensembling shows potential, achieving a 77.9% accuracy on the FakeAudio-FakeVideo class, but falls short for the RealAudio-RealVideo class.

Table 6: Classification performance of multi-modal deepfake detection methods on set A.

Models	RA-RV	RA-FV	FA-RV	FA-FV
MRDF	0.775	0.446	0.827	0.871
FACTOR	0.157	0.352	0.369	0.413
Unimodal Ensembling	0.208	0.887	0.359	0.779

To assess the impact of noise and neural compression, we evaluate baseline models trained on Sets A and B, testing them on the corrupted version of Set C (as shown in Table 1 of the Appendix). Here we observe a significant performance degradation. Additionally, we investigate different compression levels for deepfake detection models, individually training and testing them on raw, C23, and C40 compressions of Set A. From Table 2 (Appendix), DSP-FWA consistently performs well across most combinations. While there's a drop in performance when models are trained on higher quality and tested on compressed samples, those trained on C23 and C40 exhibit better generalization for both raw and C40 samples.

Protocol 2 - Zero-Day Attack Generalization: We evaluate the deepfake detection models on the challenging setting of zero-day attack detection. In this, each model is trained on training data from one set of the ILLUSION dataset and is tested on the test data of the other set. For evaluation on unseen attack setting, we train each model on train data of set A and test its performance on test data of set B, and vice-versa. The performance achieved is reported in Table 4 and Table 5. We clearly observe that all the visual and audio detection models, when trained on set A data and tested on set B data, consistently achieve random performance. The same observation is made for both visual and audio detection models when training data of set B is trained and evaluated on test data of set A. From this, we infer that the artifacts introduced in identity-swaps deepfakes and completely synthetic deepfakes are completely different. Due to this, the detection models trained on one is unable to generalize on the other. With this, we note that the proposed ILLUSION dataset will provide the researchers with a diverse range of deepfakes to capture variety of artifacts in training for better generalizability in real-world deployment.

Protocol 3 - Generalization on Real-World Deepfake Media: Since set C contains only visual deepfake media, we train the visual deepfake detection models in three different settings and report the accuracy on set C in Table 4. First, the models are trained on the image and video data of set A and then tested on set C. Then, the models are trained on set B and tested on set C. Finally, we train the models on a combination of visual data from sets A and B and test it on set C. We observe that models perform better on set C when trained on set B than when trained on set A. This behavior is observed because most samples are synthetically generated using text-to-image models like MidJourney and ArtGuru. Whereas identity swaps are very few. We also observe a slight increase in performance on set C when the detection models are trained on a combination of set A and set B.

Table 7: Classification accuracy of audio deepfake detection models (trained on Sets A and B) tested on Set D.1. Set D comprises audio samples from more than 26 languages.

Models	EER	AUC
SSLModel	0.578	0.397
Conformer	0.523	0.488
AASIST	0.506	0.471
RawGAT_ST	0.571	0.402

We further perform a comprehensive evaluation of audio and multi-modal detection models on Set D, with detailed results presented in Tables 7 and 8, respecctively. The samples in Set D were assessed using audio detection models for both subsets D.1 and D.2. For Set D.1, we employed audio detection models pretrained on a combination of Set A and Set B from the ILLUSION dataset. As shown in Table 7, all architectures encountered significant challenges in generalization, with the conformer model achieving the highest AUC of 0.488. For Set D.2, we utilized MRDF, FACTOR, and Unimodal Ensembling for multi-modal baselining and report the performance in Table 8. It is evident that model performance drastically declines when evaluated on Set D, which involves multi-modal real-world fakes. Nonetheless, FACTOR outperformed Unimodal Ensembling.

These results highlight the formidable challenge posed by the multi-lingual and multi-modal nature of Set D, revealing that current state-of-the-art algorithms struggle to generalize to real-world deepfakes.

Table 8: Classification performance of multi-modal deepfake detection methods on set D.2 of the ILLUSION dataset.

Models	RA-RV	RA-FV	FA-RV	FA-FV
MRDF	0.31	0.10	0.15	0.09
FACTOR	0.35	0.12	0.20	0.13
Unimodal Ensembling	0.25	0.09	0.11	0.05

Our observations indicate that state-of-the-art detection models face significant difficulties when confronted with the complexity of multi-modal and multi-lingual deepfakes prevalent on social media platforms. These findings highlight the challenges and limitations these models encounter in adapting to the diverse and evolving nature of deepfake content.

Table 9: Model attribution on Set A

Attribute	Models	Accuracy	AUC
Video	F3Net	0.923	0.933
	DSP-FWA	0.987	1.000
	MesoInceptionNet	0.444	0.620
	Xception	0.880	0.832
Audio	RawGAT-ST	0.941	0.995
	AASIST	0.957	0.998
	Conformer	0.967	0.999
	SSLModel	0.959	0.998

Protocol 4 - Performance on Model Attribution: Different generation techniques are shown to introduce unique nuances in the generated deepfakes (Wang et al., 2020a; Frank et al., 2020; Wang et al., 2023b). From Protocol 2, we observed that the deepfake artifacts introduced in identity-swap deepfakes are very different from those of completely synthetic deepfake media. In this experiment, we explore the performance of the detection models for the identification of source generation technique. In Table 9, the performance of all the deepfake detection models is reported for the identification of the source generation technique. We observe that for visual models, all the techniques except MesoInceptionNet are successfully able to identify the source of the identity-swap deepfakes with DSP-FWA achieving a near-perfect accuracy. A similar trend is observed for audio models, where Conformer performs the best.

For the model attribution in set B, detection models are evaluated separately for each modality. We report the performance in Table 10. For attribution in text-to-image (including synthetic faces samples) and text-to-video models, DSP-FWA consistently achieves the highest performance with an AUC of 97.8% and 99.9%, respectively. Similarly, for attribution in text-to-audio data, all the detection models are successfully able to identify the source of the generation model with comparable performance. From these observations, we note that each generative model introduces unique signatures in their generated output. The detection models pick these signatures for a near-perfect performance on model attribution task.

6 DISCUSSION AND CONCLUSION

Table 10: Model attribution on Set B

Attribute	Models	Accuracy	AUC
Images	F3Net	0.878	0.971
	DSP-FWA	0.911	0.978
	MesoInceptionNet	0.499	0.822
	Xception	0.889	0.972
Video	F3Net	0.994	0.999
	DSP-FWA	0.998	0.999
	MesoInceptionNet	0.909	0.994
	Xception	0.996	0.999
Audio	RawGAT-ST	0.991	0.999
	AASIST	0.993	0.999
	Conformer	0.989	0.998
	SSLModel	0.989	0.999

In this paper, we introduce the ILLUSION dataset, a significant step towards a comprehensive, multi-modal deepfake dataset. Created using 28 state-of-the-art generative models, ILLUSION provides diverse AI-generated content across image, audio, and video modalities and includes both curated real-world deepfakes and synthetic media. This design enables models trained on ILLUSION to learn features that extend beyond synthetic artifacts, enhancing cross-domain generalization, particularly in multi-lingual and noisy settings. Preliminary results show that detection models trained on ILLUSION outperform those trained on existing datasets when evaluated on unseen generative techniques and real-world forgeries. Designed to aid the development of robust, multi-modal, multi-lingual detection systems, our analysis of the ILLUSION dataset reveals several key insights:

Multi-Modal Deepfake Detection: The high performance of models like DSP-FWA and ASSIST on both visual and audio data suggests that current models are effective at detecting deepfakes when trained on data from same distribution. However, the disparity in performance between identity swaps and completely synthetic data indicates that models may be learning to identify artifacts specific to the generation method rather than generalizable features of deepfakes.

Zero-Day Attack Generalization: The significant drop in performance when models trained on one set are tested on another accentuates the challenge of zero-day attack detection. This suggests that models are currently not robust against deepfakes generated by unfamiliar methods, highlighting the need for diverse datasets like ILLUSION .

Generalization on Real-World Multi-Lingual Deepfake Media: The subpar performance of models trained on identity-swap and synthetic data, when tested on real-world deepfakes across various languages, depicts the necessity of a curated, multi-lingual deepfake dataset. Such a dataset is crucial for enabling models to effectively generalize to the diverse deepfakes encountered in the wild.

Model Attribution: The ability of models to identify the source generation technique with high accuracy demonstrates that generative models leave distinct signatures in their outputs. This could have implications for the traceability of deepfakes and the accountability of generative model creators.

The ILLUSION dataset focuses on addressing deepfake detection challenges through specialized generative AI techniques, while acknowledging that generalized forgery methods, such as digital watermarking, image optimization, and Photoshop-based manipulations, represent another dimension of media forensics. These generalized methods, often easier for deep-learning-based detectors to identify, differ fundamentally from generative deepfake techniques and would require additional design considerations to ensure dataset consistency. Further, despite its large scale, ILLUSION prioritizes quality and diversity over size, with each subset curated for distinct purposes, such as evaluating generalizability or robustness to compression artifacts. By incorporating 28 distinct generative methods and multi-modal, multi-lingual, and real-world samples, the dataset minimizes redundancy, ensuring relevance and providing valuable insights into detection performance across diverse conditions. Future extensions of ILLUSION will explore the integration of generalized forgery methods to further broaden its scope and utility.

7 BROADER IMPACT

Our analysis estimates that approximately 245 kg CO_2-equivalent was emitted during the creation of this dataset[11]. Despite this environmental impact, the societal benefits are significant. ILLUSION offers a valuable resource for researchers to explore detection methods across diverse types of fake

---

[11] `https://mlco2.github.io/impact/`

media. Additionally, its balanced representation of gender and skin tone promotes fairness in the development and evaluation of detection techniques. As a comprehensive multi-modal, multi-lingual deepfake dataset, ILLUSION is instrumental in the global fight against misinformation.

8 REPRODUCIBILITY STATEMENT

To promote reproducibility, we make the code, trained models, and dataset publicly available. The codebase can currently be viewed at Anonymized Repository.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7. IEEE, 2018.

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, pp. 38–45, 2019.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–10. IEEE, 2022a.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–10. IEEE, 2022b.

Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset, 2023.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2709–2720. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/casanova22a.html`.

Yiqun Chen and James Y Zou. Twigma: A dataset of ai-generated images with metadata from twitter. *Advances in Neural Information Processing Systems*, 36, 2024.

Saheb Chhabra, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Low quality deepfake detection via unseen artifacts. *IEEE Transactions on Artificial Intelligence*, 2023.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. Diff-HierVC: Diffusion-based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-shot Speaker Adaptation. In *Proc. INTERSPEECH 2023*, pp. 2283–2287, 2023. doi: 10.21437/Interspeech.2023-817.

François Chollet. Xception: Deep learning with depthwise separable convolutions, 2017.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Gölge Eren and The Coqui TTS Team. Coqui TTS, January 2021. URL `https://github.com/coqui-ai/TTS`.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Gaurav Nemade Iulia Turc. Midjourney user prompts amp; generated images (250k), 2022. URL `https://www.kaggle.com/ds/2349267`.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2889–2898, 2020.

Lea Schönherr Joel Frank. Wavefake: A data set to facilitate audio deepfake detection, 2021.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6367–6371. IEEE, 2022.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

Hasam Khalid and Simon S. Woo. OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2794–2803, 2020. doi: 10.1109/CVPRW50498.2020.00336.

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

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\’echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*, 2022.

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

Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10744–10753, October 2021.

Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5060–5069, 2019a.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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

Yuezun Li and Siwei Lyu. Dsp-fwa: Dual spatial pyramid for exposing face warp artifacts in deepfake videos. *Retrieved December*, 18:2019, 2019.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Songxiang Liu, Dan Su, and Dong Yu. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv preprint arXiv:2201.11972*, 2022.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Aakash Varma Nadimpalli and Ajita Rattani. Gbdf: gender balanced deepfake dataset towards fair deepfake detection. In *International Conference on Pattern Recognition*, pp. 320–337. Springer, 2022.

Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Deephy: On deepfake phylogeny. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10. IEEE, 2022.

Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Df-platter: Multi-face heterogeneous deepfake dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9739–9748, 2023.

Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme, 2022.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *arXiv*, 2023.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

Tal Reiss, Bar Cavia, and Yedid Hoshen. Detecting deepfakes without seeing any. *arXiv preprint arXiv:2311.01458*, 2023.

Joseph Robinson. Balanced faces in the wild, 2022. URL `https://dx.doi.org/10.21227/nmsj-df12`.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019.

s0md3v. Roop. https://github.com/s0md3v/roop, 2023.

Davide Salvi, Brian Hosler, Paolo Bestagini, Matthew C. Stamm, and Stefano Tubaro. Timit-tts: a text-to-speech dataset for multimodal synthetic media detection, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv preprint arXiv:2107.12710*, 2021.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233*, 2022.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020a.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023b.

Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12321–12328, 2020.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Ying Xu, Philipp Terhörst, Kiran Raja, and Marius Pedersen. A comprehensive analysis of ai biases in deepfake detection with massively annotated databases. *arXiv preprint arXiv:2208.05845*, 2022a.

Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu, Junyu Han, Jingtuo Liu, and Errui Ding. Mobilefaceswap: A lightweight framework for video face swapping, 2022b.

Massimiliano; Sahidullah Md; Delgado Héctor; Wang Xin; Evans Nicolas; Kinnunen Tomi; Lee Kong Aik; Vestman Ville; Nautsch Andreas. Yamagishi, Junichi; Todisco. Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *CVPR*, 2023.

Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 440–445. IEEE, 2023.

Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8568–8577, 2023.

Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5778–5788, 2021.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2382–2390, 2020.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*, 2024.

Heqing Zou, Meng Shen, Yuchen Hu, Chen Chen, Eng Siong Chng, and Deepu Rajan. Cross-modality and within-modality regularization for audio-visual deepfake detection. *arXiv preprint arXiv:2401.05746*, 2024.