Large-Scale Chinese Benchmark for Face Video Anti-Spoofing Detection

Authors: Bei Yijun, Lou Hengrui, Gao Kewei, Song Jie, Wang Rui, Jin Canghong, Song Mingli, Feng Zunlei, Feng Zunlei

Date: 2024-01-22T00:00:00+00:00

Abstract

With the rapid development of AIGC technology, realistic forged face videos can already deceive human visual perception. Consequently, numerous face anti-spoofing detection algorithms have been proposed for detecting forged face videos. However, effectively evaluating the effectiveness and applicability of these detection algorithms remains challenging. To promote the quantitative evaluation of face anti-spoofing detection performance and the iterative development of anti-spoofing detection technology, this paper proposes a large-scale Chinese data benchmark for face video anti-spoofing detection and releases the world's first CHN-DF Chinese dataset (https://github.com/HengruiLou/CHN-DF), filling the gap in large-scale Chinese data for face video anti-spoofing datasets. This paper elaborates on the construction process of the CHN-DF dataset and the Chinese data evaluation benchmark, and validates through experiments the complexity of the CHN-DF dataset and its proximity to real-world scenarios. We expect this benchmark to assist researchers in constructing more practical and effective face video anti-spoofing detection models, advancing technological development in the anti-spoofing detection field. Additionally, this paper identifies the challenges confronting Chinese face video anti-spoofing benchmark datasets and anti-spoofing detection technology, proposes potential future research directions, and provides valuable insights for advancing the development of face video anti-spoofing detection technology.

Full Text

Large-Scale Chinese Data Benchmark for Face Video Anti-Forgery Detection

Yi-Jun Bei[1,2], Heng-Rui Lou[1], Ke-Wei Gao[1], Jie Song[1,2], Rui Wang[3], Cang-Hong Jin[4], Ming-Li Song[2], Zun-Lei Feng[1,2]

[1]College of Software Technology, Zhejiang University, Ningbo 315003, China
[2]College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China
[3]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[4]College of Computer and Computer Science, Hangzhou City University, Hangzhou 310015, China

Corresponding Author: Zun-Lei Feng, E-mail: zunleifeng@zju.edu.cn

Abstract

With the rapid development of AIGC (Artificial Intelligence Generated Content) technology, hyper-realistic forged facial videos have become capable of deceiving human visual perception. Consequently, numerous facial anti-forgery detection algorithms have been proposed to identify these fake videos. However, effectively evaluating the efficacy and applicability of these detection algorithms remains a substantial challenge. To promote quantitative assessment of facial anti-forgery detection performance and iterative advancement of anti-forgery technologies, this paper introduces a large-scale Chinese data benchmark for facial video anti-forgery detection and releases the world's first CHN-DF Chinese dataset (https://github.com/HengruiLou/CHN-DF), filling the gap in large-scale Chinese data for facial video anti-forgery datasets. We detail the construction process of the CHN-DF dataset and the Chinese data evaluation benchmark, and validate the dataset's complexity and realism through extensive experiments. We hope this benchmark will assist researchers in building more practical and effective facial video anti-forgery detection models, thereby advancing the field. Additionally, this paper addresses the challenges facing Chinese facial video anti-forgery detection benchmark datasets and detection technologies, proposing future research directions to provide valuable insights for advancing facial video anti-forgery detection technology.

Keywords: Deep Learning; Deepfakes; Fake Video; Multimodal Anti-Forgery Detection

1. Related Work on Video Deepfake Datasets

The transformation of video content generation driven by AIGC development has intensified the urgency of detecting facial forgeries. In recent years, numerous researchers from academia and industry have dedicated themselves to creating facial video anti-forgery detection datasets, with several open-sourced datasets advancing research in this domain. This section surveys the current state of facial video anti-forgery detection datasets (summarized in Table 1).

Existing facial video anti-forgery detection datasets fall into two main categories. The first category employs unimodal visual forgery methods that modify or swap facial features to achieve face forgery effects. The second category combines visual and auditory forgery techniques, applying multimodal modifications to

visual or auditory feature information in authentic videos to achieve complex video forgeries. These multimodal approaches offer diverse forgery angles and methods that better reflect real-world malicious video manipulation scenarios, representing the development trend for video deepfake datasets. However, they require diverse and complex forgery techniques, resulting in scarce data samples.

1.1 Unimodal Visual Facial Video Anti-Forgery Detection Datasets

- **UADFV** [**?**]: Released in 2018 by researchers at the State University of New York, UADFV was the first dataset for facial video anti-forgery detection. It contains 98 videos total—49 authentic videos collected from YouTube and 49 fake videos generated using the FakeApp application [**?**]. Videos average 11.14 seconds in length with a resolution of 294$×$500 pixels. As an early dataset, UADFV suffers from limited quantity and quality, with obvious facial distortions and abnormal movements in the single-method FakeApp-generated videos that make detection relatively easy.
- **DeepfakeTIMIT** [**?**]: Also introduced in 2018, this dataset targets deepfake detection with authentic data from 640 videos of 32 speakers. Each speaker's video set contains 10 high-resolution DeepFake-TIMIT-HQ videos and 10 low-resolution DeepFake-TIMIT-LQ videos. Fake videos were created by swapping facial information between speakers. However, due to early video forgery limitations, generated videos are only 4 seconds long and often blurry.
- **FF++** [**?**]: This dataset employs four forgery methods—Deepfake [**?**], Face2face [**?**], Faceswap [**?**], and NeuralTextures [**?**]—making it the first to include both deep learning-based and computer graphics-based forgery techniques. It contains 1,000 authentic videos from YouTube and 4,000 fake videos synthesized using computer graphics and two deep learning methods. Additionally, the dataset is divided into two quality levels (uncompressed and H264 compression) to evaluate detection performance on compressed versus uncompressed videos. However, FF++ lacks sufficient size and diversity for optimal training of high-performance neural architectures with numerous parameters.
- **Celeb-DF** [**?**]: Addressing quality issues and coarse manipulation artifacts in UADFV, FF++, and DeepfakeTIMIT, Celeb-DF improved video forgery methods to provide higher quality videos. Authentic videos were sourced from 590 videos of 59 YouTube speakers, with 5,639 fake videos generated using improved deepfake technology. However, the dataset still suffers from single-method forgery limitations, making it unsuitable for real-world challenges.
- **DeeperForensics** [**?**]: Authentic videos were recorded from 100 paid actors, with 1,000 target videos from FF++ used for face-swapping forgery. By swapping each source identity with 10 target videos, 1,000 fake videos

were synthesized. Rather than employing additional synthesis methods, DeeperForensics applied seven perturbation methods for data augmentation on both real and fake videos, creating 50,000 real and 10,000 fake videos. While significantly larger and more diverse than early datasets, DeeperForensics has not been extensively evaluated against current face forgery techniques, leaving its academic utility not fully realized.

- **WildDeepfake** [**?**]: To address the lack of content diversity and low-quality video sources in early datasets, WildDeepfake collected real and deepfake samples from the internet, including facial action sequences extracted from videos. After manually removing videos without corresponding real faces, the dataset contains 3,805 real and 3,509 fake videos. The visual effects better match real-life scenarios, but insufficient data volume limits training of high-performance neural network architectures.
- **ForgeryNet** [**?**]: Currently the largest-scale visual-based facial video anti-forgery detection dataset, ForgeryNet proposes multiple tasks including temporal forgery localization and spatial forgery localization. It employs eight deepfake methods to generate 121,617 fake videos, with a total of 221,247 videos featuring rich annotations.

1.2 Multimodal Visual-Auditory Facial Video Anti-Forgery Detection Datasets

- **DFDC** [**?**]: The first dataset to include forged audio in videos, initially released as the dataset for Facebook's DFDC competition with 5,250 videos. After data supplementation, it reached 23,654 real videos and 104,500 fake videos. To ensure diversity, authentic video sources were captured in various environmental settings, with fake videos generated by eight different methods. Auditory modality only involved audio swapping without audio forgery methods. Labels contain only real/fake categories without distinguishing visual versus auditory forgeries.
- **KoDF** [**?**]: Currently the largest publicly available multimodal facial video anti-forgery detection dataset, containing 175,776 fake videos forged by six methods and 62,166 real videos. The 403 speakers are predominantly Korean, representing the first effort to balance the underrepresentation of Asian populations in existing anti-forgery datasets. However, KoDF only synchronizes audio with lip movements without using deep voice forgery methods like voice cloning or voice conversion.
- **FakeAVCeleb** [**?**]: The first facial video anti-forgery detection dataset to simultaneously include forged video and forged audio, commonly used as a benchmark for multimodal facial video anti-forgery detection. It selected 500 real videos from the VoxCeleb2 dataset, used Faceswap, DeepFaceLab [**?**], and FSGAN for facial forgery, SV2TTS [**?**] for audio forgery, and Wav2Lip for audio-lip synchronization, generating 19,500 fake videos.

2. The CHN-DF Facial Video Anti-Forgery Detection Dataset

CHN-DF is the first large-scale Chinese dataset for facial video anti-forgery detection, containing information from both visual and auditory modalities. This section first describes authentic video collection and fake video generation for CHN-DF, then details the dataset's fundamental attributes.

2.1 Authentic Videos To ensure scene diversity and content complexity, CHN-DF authentic videos are sourced from CN-CVS, the largest publicly available Chinese audio-visual multimodal dataset, and CMLR, a Chinese lip-reading dataset. CN-CVS contains over 2,500 speakers with more than 200,000 clips totaling over 300 hours. CHN-DF selected videos from 2,529 speakers in the Speech portion of CN-CVS. The CMLR dataset comprises news broadcast videos from June 2009 to June 2018, containing 102,076 videos presented by 11 anchors. CHN-DF filtered the CMLR dataset to balance video quantities across speakers, selecting approximately 20,000 videos. Consequently, CHN-DF authentic video volume reaches 213,187, exceeding authentic video counts in existing public facial video anti-forgery detection datasets, with 2,540 total speakers.

Since CN-CVS and CMLR videos are already cropped to facial regions (CMLR using HOG-based face detection and open-source platforms for face recognition and alignment; CN-CVS using the dlib toolkit for face detection with videos lacking faces or containing multiple faces removed), CHN-DF requires no additional face detection or localization. As videos are partitioned by speaker identity, training, validation, and test sets contain no overlapping speakers, giving CHN-DF high extensibility. New speakers' authentic and fake videos can be easily added while maintaining mutual independence among splits.

2.2 Fake Videos CHN-DF fake videos employ seven state-of-the-art deepfake methods covering mainstream approaches: Mockingbird, coqui-TTS, Wav2Lip, SimSwap, FOMM, Motion-cos, and FSGAN. SimSwap and FSGAN are face-swapping methods; FOMM and Motion-cos are face reenactment methods; Mockingbird and coqui-TTS are voice cloning methods; Wav2Lip is a lip-synchronization method. Figure 1 [Figure 1: see original paper] shows examples generated by selected visual forgery methods, with each row displaying frames created sequentially by Wav2Lip, SimSwap, FOMM, Motion-cos, and FSGAN. Figure 2 [Figure 2: see original paper] illustrates the distribution of fake videos by method. Although quantities vary due to manual screening based on forgery quality, CHN-DF maintains relative balance across methods. The "others" category refers to videos generated by replacing source video audio with audio from other videos within the same subset (training, validation, or test).

- **Mockingbird**: Used for real-time Chinese voice cloning, Mockingbird [**?**] synthesizes fake audio from different speakers' audio information. Building upon SV2TTS, Mockingbird introduces Chinese training datasets (ai-

datatang_{200zh}, magicdata, aishell3) to train the speech synthesis system, processing speech to extract speaker voice timbre vectors (Speaker Encode), then completing Chinese speech synthesis via synthesizer and vocoder.

- **coqui-TTS**: A low-resource, zero-shot text-to-speech model [**?**] capable of synthesizing multiple languages including Chinese. It provides various text-to-speech models (Tacotron [**?**], Tacotron2 [**?**], Glow-TTS [**?**]) and vocoder models (MelGAN [**?**], Multiband-MelGAN [**?**], GAN-TTS [**?**]), enabling high-quality speech output for complex text-to-speech conversion tasks.
- **Wav2Lip**: A GAN-based lip-motion transfer algorithm [**?**] that generates lip-synchronized videos from images and target speech or directly converts dynamic videos' lip shapes to match input audio. Using a pre-trained lip-sync detector, the model learns lip movements from audio, capturing temporal context through five consecutive face frames with corresponding speech content as input.
- **SimSwap**: This model employs an Identity Injection Module (IIM) [**?**] that transfers source image facial identity information to target videos at the feature level, using weak feature matching loss to implicitly preserve facial attributes, enabling universal and high-fidelity face swapping.
- **FOMM**: A self-supervised model that disentangles appearance and motion information [**?**], consisting of motion estimation and image generation modules. By observing frame pairs from the same video, the model encodes motion as keypoint displacements and local affine transformations specific to motion, reconstructing training videos from learned motion feature maps. During application, the model pairs source images with target video frames to animate source objects, generating fake videos of the source face.
- **Motion-cos**: A self-supervised deep learning method for part segmentation [**?**] that extracts keypoint information from source face images and performs frame-by-frame forgery on target videos based on sub-component feature maps, enabling regional face swapping. Motion-cos provides pre-trained models for five-, ten-, and fifteen-segment face partitioning; CHN-DF adopts the fifteen-segment model for fine-grained face swapping.
- **FSGAN**: An adversarial generative network-based face swapping model [**?**] that performs face swapping and reenactment using target and source videos. The model redraws source faces according to target pose and expression, segments them into two facial regions, fills missing parts, and blends the complete face with the target. During reenactment, Delaunay triangulation selects multiple source frames best matching the target face, using barycentric coordinates for weighted averaging of reenactment results without requiring extensive adjustment for each new source video. CHN-DF employs FSGAN's face swapping technique.

2.3 Dataset Description

2.3.1 Category Description Using these deepfake methods, CHN-DF is categorized into four types based on visual and auditory modalities: Real Visual-Real Auditory (VRAR), Real Visual-Fake Auditory (VRAF), Fake Visual-Real Auditory (VFAR), and Fake Visual-Fake Auditory (VFAF).

Table 2 CHN-DF Visual-Auditory Forgery Combinations and Methods

CHN-DF Authentic Visual Source (VR)	Fake Visual Generation (VF)	Authentic Auditory Source (AR)	Fake Auditory Generation (AF)
SimSwap, FOMM, Motion-cos, FSGAN	Mockingbird, coqui-TTS	Wav2Lip	VF

(1) **Real Visual-Real Auditory (VRAR)**: VRAR data originates from CN-CVS and CMLR. From CN-CVS/Speech, we selected videos of 2,529 speakers featuring diverse speakers and complex, variable environments that reflect real-life conversation scenarios. From CMLR, we filtered approximately 20,000 anchor presentation videos of 11 hosts. Speaker identities were numbered, yielding 213,187 VRAR videos.

(2) **Real Visual-Fake Auditory (VRAF)**: VRAF maintains authentic visuals while forging audio. As shown in Table 2, Mockingbird and coqui-TTS generate cloned fake audio. Specifically, source video speakers' text statements and other speakers' audio serve as model inputs to generate fake audio cloned from others' voices, which is then merged with source videos. This deepfake category simulates identity fraud through voice imitation, useful for training defenses against voice spoofing attacks. VRAF contains 63,070 videos.

(3) **Fake Visual-Real Auditory (VFAR)**: VFAR forges faces while preserving authentic audio. Face forgery employs face swapping (SimSwap, FSGAN) and face reenactment (FOMM, Motion-cos). Face swapping replaces source video faces with other speakers' faces, while reenactment applies other speakers' facial actions to source video faces. Merging forged video with source audio yields VFAR videos. This category trains defenses against identity fraud where attackers modify facial actions or swap faces to create non-existent video footage. VFAR contains 88,888 videos.

(4) **Fake Visual-Fake Auditory (VFAF)**: VFAF combines both face and audio forgery, integrating VRAF and VFAR methods with Wav2Lip (Table 2). It includes three forgery approaches: (1) merging temporally similar forged audio and video; (2) merging then applying Wav2Lip for lip-sync ("lip matching"); (3) applying Wav2Lip to VRAF videos to alter lip movements. VFAF integrates VRAF and VFAR categories, reflecting complex

real-world scenarios with simultaneous audio-visual forgery. VFAF contains 60,942 videos.

Notably, "other videos" mentioned in VRAF, VFAR, and VFAF forgery processes are always from the same subset (training, validation, or test) as source videos, ensuring mutual independence among splits.

2.3.2 Dataset Attributes CHN-DF contains 426,087 facial videos from 2,540 speakers, including 213,187 authentic videos and 212,900 fake videos, achieving balanced positive-negative samples. Fake video categories (VRAF, VFAR, VFAF) contain approximately 63,070, 88,888, and 60,942 videos respectively.

Videos are partitioned by speaker identity into training (350,679 videos from 1,778 speakers), validation (22,685 videos from 254 speakers), and test sets (52,723 videos from 508 speakers) at a 7:1:2 ratio. Figure 3 [Figure 3: see original paper] shows CHN-DF video duration distribution, ranging from 0.36 to 355.58 seconds (average 5.12 seconds), reflecting real-world video length variability. Durations concentrate at 0-20 seconds, with 98.75% of clips under 20 seconds and 99.94% under 50 seconds.

3. CHN-DF Benchmark Evaluation

The ultimate goal of creating facial video anti-forgery detection datasets is to develop models that perform well across various deepfake types and methods. Model performance is evaluated through multiple quantitative metrics on these datasets. This section describes CHN-DF benchmark evaluation methods and metrics, presenting comprehensive performance assessments using eight state-of-the-art multimodal facial video anti-forgery detection methods to demonstrate CHN-DF's complexity and realism. We compare results with the recently published multimodal FakeAVCeleb dataset, selected because it is currently the only publicly available multimodal dataset with detailed audio-video forgery annotations and rich forgery methods, making it a widely accepted benchmark in multimodal facial video anti-forgery detection [**?**].

3.1 Evaluation Methods Given CHN-DF's visual and auditory modalities, we evaluate both ensemble methods integrating single-modal detection results and multimodal facial video anti-forgery detection models.

3.1.1 Ensemble Methods

(1) **Meso-4**: A four-layer convolutional network proposed by Afchar et al. [**?**] for face forgery detection based on mid-level image noise information. This approach effectively addresses issues of diminished image noise and difficulty distinguishing fake video frames through high-level semantic features. Its shallow architecture enhances sensitivity to medium and large-scale features, improving facial feature detection but limiting capture of deeper, subtler features.

(2) **MesoInception-4**: Also proposed by Afchar et al. [**?**], inspired by InceptionNet [**?**], this model improves Meso-4 by replacing the first convolutional layer with InceptionNet modules to capture multi-scale features more effectively, though it still shares shallow architecture limitations.

(3) **Xception**: A convolutional neural network architecture based entirely on depthwise separable convolution layers [**?**], derived from simplifying channel and spatial correlation decoupling. It efficiently extracts complex features from images and video frames, though its complex structure may complicate training and tuning.

3.1.2 Multimodal Methods

(1) **Multimodal-2**: An open-source multimodal model for movie genre prediction [**?**], comprising a CNN block for movie posters (visual modality), an LSTM block for movie genres (text modality), and a feedforward network for classification combining both outputs. For fake video detection, it uses the CNN block to analyze subtle differences in video frames and the LSTM block to process audio temporal information, effectively capturing inconsistencies in forged videos.

(2) **CDCN**: A central difference convolutional network for face anti-spoofing [**?**] that employs three-layer fused features (low, medium, high) to predict grayscale facial depth. Compared to traditional CNNs, CDCN effectively extracts subtle local features like skin texture and expression details to capture minute artifacts from forgery techniques.

(3) **MDS**: Audio-visual synchronization is difficult to forge successfully, as forged video frames often exhibit lost lip shapes or unnatural facial and lip movements. MDS [**?**] compares visual and auditory content in fake videos, detecting multimodal forgeries by quantifying cross-modal incoordination.

(4) **VFD**: VFD [**?**] focuses on matching degree between human biometric features (voice and face), leveraging intrinsic correlations for facial anti-forgery detection by learning essential facial and audio features to bring matching audio-visual pairs closer while separating mismatched ones.

(5) **AVoiD-DF**: An audio-visual joint learning method for multimodal facial video forgery detection [**?**], comprising a Temporal-Spatial Encoder (TSE), Multi-Modal Decoder (MMD), and Cross-Modal Classifier. It detects forgeries through audio-visual inconsistencies at spatiotemporal levels.

3.2 Evaluation Metrics We adopt four metrics to evaluate model performance on datasets: Accuracy, Precision, Recall, and F1-score. These combinations are chosen for their widespread use in classification and particular suitability for facial video anti-forgery detection:

(1) **Comprehensive Performance Assessment**: Accuracy measures overall correct prediction proportion, providing a holistic performance per-

spective for balanced datasets. F1-score combines Precision and Recall, offering balanced measurement between positive and negative samples, particularly valuable for imbalanced datasets.

(2) **Enhanced Focus on Anti-Forgery Scenarios**: In security applications like facial video anti-forgery detection, we prioritize true positive capture rates—how many model results are genuine positives. Precision and Recall directly reflect model performance on true positives.

(3) **Handling Class Imbalance**: Facial video anti-forgery datasets often exhibit class imbalance. F1-score is appropriate for evaluating model performance under such conditions, better reflecting positive sample classification capability.

(4) **Reflecting Dataset Quality**: For balanced datasets, Accuracy can reflect overall dataset quality. Under imbalance, F1-score more sensitively reflects model handling of minority classes, better evaluating dataset quality.

3.3 Benchmark Experiments

3.3.1 Dataset Preprocessing For training benchmark models, we preprocess both modalities separately. For visual modality, since CHN-DF source datasets (CMLR and CN-CVS) are already cropped to facial regions, no face detection is needed. We extract and store video frames separately as model inputs for visual feature extraction. For auditory modality, we first extract audio at 16kHz sampling rate and store as WAV format. We then compute Mel-Frequency Cepstral Coefficients (MFCC) features using 25ms Hann windows with 10ms shifts, obtaining 80 MFCC features per audio frame (D=80). These MFCC features are stored as three-channel images, with quantities upsampled to address the single-MFCC-image-per-video issue. These MFCC images serve as inputs for speech feature extraction to learn distinctions between real and fake speech.

3.3.2 Benchmark Experimental Setup For fair CHN-DF benchmarking, we adopt the same model parameters used for FakeAVCeleb evaluation. Specifically, each method trains for 50 iterations with EarlyStopping (patience=10), using Adam optimizer at learning rate 10^{-5}. Experiments run on a computer with Silver 4310 CPU and Nvidia A40 GPU. For ensemble methods, Hard-Voting and Soft-Voting mechanisms integrate predictions from audio and video anti-forgery models [**?**].

3.3.3 Experimental Results and Analysis **Table 3** Multimodal Anti-Forgery Method Comparison on CHN-DF Dataset

Methods	CHN-DF			FakeAVCeleb		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Meso-4(Soft-Voting)						
Meso-4(Hard-Voting)						
MesoInception-4(Soft-Voting)						
MesoInception-4(Hard-Voting)						
Xception(Soft-Voting)						
Xception(Hard-Voting)						
Multimodal-2						
AVoiD-DF						

We select FakeAVCeleb for comparison because, as described in Section 1.2, it is currently the only publicly available multimodal dataset with both deepfake audio and video. To analyze benchmark model performance and validate CHN-DF's complexity and realism, we evaluate selected methods on both CHN-DF and FakeAVCeleb. Performance results appear in Table 3, with best results between datasets bolded.

3.3.3.1 Validation of CHN-DF Complexity and Realism

Among 44 total metrics from 11 benchmark models (distinguishing Hard-Voting and Soft-Voting), 32 metrics are lower on CHN-DF than FakeAVCeleb. Benchmark model results on CHN-DF concentrate below 0.6 across all four metrics. Performance on CHN-DF is inferior to FakeAVCeleb in Precision, Recall, Accuracy, and F1-score, indicating more complex and challenging anti-forgery tasks. This validates CHN-DF's complexity and realism, making it more conducive to developing better deepfake detection methods.

3.3.3.2 Benchmark Model Performance Analysis

AVoiD-DF achieves optimal performance on both CHN-DF and FakeAVCeleb, likely due to its Multi-Modal Decoder (MMD) module for modality fusion. Unlike other multimodal methods, AVoiD-DF feeds visual and audio embedding blocks through two parallel decoder channels, each with a Bidirectional Cross-Attention (BiCroAtt) module followed by self-attention blocks and feedforward layers, enabling better information sharing and joint learning between modalities. However, AVoiD-DF's performance drops significantly on CHN-DF, possibly because its audio-visual joint learning approach struggles with Fake Visual-Fake Auditory (VFAF) cases. Wav2Lip's lip conversion disrupts intrinsic face-audio correlations, and CHN-DF employs more complex forgery methods than FakeAVCeleb, yielding lower metrics on VFAF data.

MesoInception-4 performs best among ensemble methods, likely because it uses Inception variants to focus on artifacts from resolution inconsistencies between warped face regions and surroundings during affine transformations—a common limitation in forgery methods. However, MesoInception-4 struggles

with FOMM and Motion-cos reenactment videos, as reenactment preserves target identity while applying source features, producing different artifacts than face swapping. This limitation yields lower metrics on FOMM and Motion-cos generated videos.

Multimodal-2 and **Xception** show poor performance on both datasets (CHN-DF metrics below 0.52), possibly because these general computer vision classification models, while effective across various tasks, suffer from domain gaps between pretrained weights and this specific task. Facial video anti-forgery involves richer information including expressions and poses that general models may not capture effectively.

Multimodal methods outperform ensemble methods, likely because they exploit visual-auditory correlation and consistency information. Forging cross-modal correlations is more difficult than single-modal forgery, making such inconsistencies more detectable and providing clearer detection features.

3.3.4 Cross-Dataset Forgery Detection Comparison To evaluate CHN-DF quality and benchmark model generalization, we conduct cross-dataset experiments training on FakeAVCeleb and testing on CHN-DF. Training on FakeAVCeleb enables models to learn distributions of facial forgery videos, while testing on CHN-DF reveals performance on differently distributed data, validating robustness and generalization. Comparing results with CHN-DF-trained models also assesses CHN-DF dataset quality.

Table 4 Cross-Dataset Forgery Detection Comparison

Methods	Precision	Recall	F1-score
Meso-4(Soft-Voting)			
Meso-4(Hard-Voting)			
MesoInception-4(Soft-Voting)			
MesoInception-4(Hard-Voting)			
Xception(Soft-Voting)			
Xception(Hard-Voting)			
Multimodal-2			
AVoiD-DF			

All 11 models show significantly degraded metrics in this cross-dataset task, indicating CHN-DF presents more complex and challenging forgery data, further validating its complexity and realism. Table 4 shows cross-dataset results. Compared with Table 3, performance drops significantly due to different data sources. **MesoInception-4** shows the most dramatic degradation, likely because FakeAVCeleb lacks reenactment-based fake videos for training, exacerbating limitations in artifact-based detection. **VFD** achieves optimal performance despite degradation, likely due to its fine-tuning mechanism based on pretrained models enabling rapid adaptation. **Multimodal-2**, **Xception**, and **MDS** show

smaller degradation, possibly because their general classification architecture provides good generalization despite limited video data suitability.

4. Challenges and Future Directions

Forgery and detection exist in a complex adversarial yet mutually supportive relationship. While facial video anti-forgery detection has advanced considerably to combat evolving forgery techniques, AIGC's rapid development now produces highly realistic images and videos, significantly impacting detection capabilities. Detection technology currently lags far behind forgery technology, making accurate fake video detection extremely challenging. Realistic facial video anti-forgery detection datasets are thus essential for developing more effective detection models. Moreover, existing datasets predominantly feature Western populations, with international Chinese anti-forgery datasets notably absent. Building a large-scale Chinese benchmark for facial video anti-forgery detection is crucial for advancing deepfake detection technology.

4.1 Benchmark Dataset Construction Limitations Our proposed large-scale Chinese benchmark, while pioneering, still faces challenges in authenticity, diversity, accuracy, and adversarial robustness. Addressing these to build higher-quality benchmarks is vital for advancing deepfake detection technology. Key limitations include:

(1) **Deepfake Technology Limitations**: While AIGC enables more realistic image and video generation, current techniques still exhibit issues in long video generation: (i) transient facial flickering during speech; (ii) blurred forged facial region edges; (iii) over-smoothed facial textures lacking detail; (iv) unnatural head pose movements; (v) absence of facial occlusions like glasses or lighting effects; (vi) sensitivity to body pose or skin color consistency changes causing identity leakage; (vii) unnatural emotional expression and breathing patterns. These forgery artifacts are features detection models must learn, but overemphasis causes overfitting and poor real-world robustness. Such artifacts also interfere with benchmark accuracy and objectivity, though currently unavoidable given limitations in ensuring naturalness, fluency, and continuity. To mitigate bias, we manually screen videos with overly obvious artifacts to reduce low-quality interference in quantitative evaluation.

(2) **Lack of Speech Diversity**: Current benchmarks lack diverse speech data, particularly in emotional expression, limiting comprehensive emotion detection testing. Collecting diverse cultural background speech data poses significant challenges, especially for Chinese—the world's most spoken language with multiple dialects and accents varying across regions and social groups. This deficiency may cause benchmarks to underperform on specific accents or speech styles. Building diverse, personalized speech samples is a primary future direction.

(3) **Label Inaccuracy**: Effective benchmarks require realistic datasets with accurate labels. However, large-scale annotation may introduce subjective and inconsistent labeling, particularly for high-quality AIGC forgeries that consume time yet yield inaccurate labels. Fine-grained labels require annotators with deep forgery technique expertise, which they may lack. Such subjectivity and inconsistency pose challenges to label accuracy.

(4) **Vulnerability to Adversarial Attacks**: Benchmarks lack adversarial attacks. Real-world attackers consider adversarial perturbations to reduce detection effectiveness, such as adjusting lighting intensity to hinder visual feature extraction. This causes trained models to be vulnerable to adversarial attacks. Such complex scenarios are difficult to effectively incorporate during benchmark construction, leaving detection algorithms challenged in real-world deployment.

4.2 Challenges in Facial Video Anti-Forgery Detection Technology Benchmarks and detection technologies mutually promote development in an adversarial relationship. While AIGC's rapid evolution challenges existing benchmarks, detection research also faces significant obstacles:

(1) **Difficulty Detecting Large Model-Generated Content**: Early fake videos exhibited visual artifacts or audio distortion, but widespread AIGC application in video generation makes detection harder. Large language models like ChatGPT-4.0 and DALL-E for video generation [**?**], combined with diffusion models that synthesize data from pure noise by reversing Gaussian noise addition [**?**], make forgery clues difficult to capture, posing major challenges.

(2) **Difficulty Handling Complex Scene Forgeries**: Complex scene diversity increases detection complexity. Real-world detection suffers environmental interference: lighting changes alter facial shadows and highlights; camera angle changes cause shape distortion; background complexity may blur facial edges or cause merging. These factors affect authenticity and increase detection difficulty.

(3) **Poor Generalization**: While detection techniques perform well on single datasets, cross-dataset generalization remains inadequate. In real scenarios with unknown forgery methods, detection credibility using pretrained models cannot be guaranteed.

(4) **Single-Grained Detection Tasks**: Current detection focuses on video-level forgery detection. However, attackers often forge only a few frames or audio segments. Video-level detection models may miss these subtle forgeries, increasing misjudgment probability.

4.3 Future Directions Despite significant progress, the field faces numerous challenges requiring solutions. Focusing on facial video anti-forgery detection technology and benchmarking, we provide new perspectives and directions. For

benchmarks, consider objective quantification and dynamic updates; for detection technology, consider autonomous evolution mechanisms and robustness. Additionally, data privacy protection and social impact warrant consideration.

(1) **Objective Benchmark Quantification**: Current benchmarks rely on specific model performance metrics, creating evaluation angle limitations. Future benchmarks should precisely quantify multi-angle detection capabilities and even model adaptability in real scenarios.

(2) **Dynamic Benchmark Updates**: Design should account for diverse forgery categories. Regularly updating benchmarks with latest forgery techniques helps maintain relevance to complex real-world scenarios. Integrating user feedback data provides new ideas for dynamic updates. As deepfake technology evolves, establishing dynamic label update mechanisms for new techniques becomes increasingly important.

(3) **Novel Forgery Detection**: With rapid development of generative diffusion models and large models, current generation quality increasingly matches real videos. Previous anti-forgery methods targeting synthetic artifacts or blur cannot handle high-fidelity generated videos. Future research should design techniques based on differences between real and fake videos themselves, including local feature similarity in models and inference path differences.

(4) **Emphasizing Model Robustness**: Robustness is key to stability and reliability in complex real-world scenarios. Training and testing with compression and noise interference can simulate real distribution shifts, building high robustness. Adversarial sample inclusion also promotes robustness. However, while noise and adversarial samples enhance robustness, they may also degrade recognition performance. Future research should mine differences between real and fake samples from intrinsic features to build methods handling arbitrary fakes while maintaining accuracy.

(5) **Autonomous Evolution Frameworks**: Forgery and anti-forgery are mutually advancing technologies where forgery generally leads. This performance gap causes significant societal harm. Current anti-forgery design relies on researchers analyzing forgery weaknesses. Future research should leverage adversarial learning and reinforcement learning to design autonomous evolution frameworks enabling rapid adaptation to evolving forgeries.

(6) **Data Privacy Protection**: Benchmark construction and detection must consider sensitive information privacy protection. Anonymization and other privacy-preserving techniques should ensure no personal privacy leakage during model evaluation and application, respecting user privacy rights as legal frameworks mature.

(7) **Social Impact Research**: The deepfake field lacks comprehensive legal systems for precise control, such as distinguishing entertainment from

malicious content. Establishing legal frameworks to penalize malicious creators and distributors is needed [**?**]. Researching societal impacts and ethics will provide comprehensive understanding and promote sustainable development, considering social responsibility alongside technical progress.

Conclusion

In the AIGC era where verifying authenticity in facial video generation is increasingly difficult, this paper proposes a large-scale Chinese data benchmark for facial video anti-forgery detection, releasing the world's first large-scale Chinese dataset—CHN-DF—to fill the gap in Chinese data for this domain. We detail CHN-DF dataset and benchmark construction processes, conduct comparative experiments with mainstream detection methods, and analyze existing method strengths and weaknesses from benchmark performance and cross-dataset generalization perspectives. Extensive experiments validate CHN-DF's complexity and realism. We hope this Chinese benchmark will help researchers build superior facial video anti-forgery detection models and serve as a cornerstone for future research. Additionally, we identify current challenges and future directions for Chinese facial video anti-forgery datasets and benchmarks, aiming to provide new perspectives for advancing the field.

References

[1] Han Y, Li SY, Liu YX, Yan ZL, Dai YT, Philip S, Sun LC. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv preprint arXiv:2303.04226. 2023.
[2] Zhao WX, Zhou K, Li JY, Tang TY, Wang XL, Hou YP, Min YQ, Zhang BC, Zhang JJ, Dong ZC, Du YF, Yang C, Chen YS, Chen Z, Jiang JH, Ren RY, Li YF, Tang XY, Liu ZK, Liu PY, Nie JY, Wen JR. A Survey of Large Language Models. arXiv preprint arXiv:2303.18223. 2023.
[3] Nguyen HH, Yamagishi J, Echizen I. Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467.
[4] Yang X, Li Y, Lyu S. Exposing deep fakes using inconsistent head poses. In: Proc. of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: ICASSP, 2019. 8261-8265.
[5] Shao R, Wu TX, Nie LQ, Liu ZW. DeepFake-Adapter: Dual-Level Adapter for DeepFake Detection. arXiv preprint arXiv:
[6] Haliassos A, Vougioukas K, Petridis S, Pantic M. Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proc. of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: CVPR, 2021. 5037-5047.
[7] Chen HS, Rouhsedaghat M, Ghani H, Hu S, You S, Kuo CC. DefakeHop: A light-weight high-performance deepfake detector. In: Proc. of the 2021 IEEE International Conference on Multimedia and Expo. Shenzhen: ICME, 2021. 1-6.
[8] Wodajo D, Atnafu S. Deepfake video detection using convolutional vision

transformer. arXiv preprint arXiv: 2102.11126. 2021.
[9] Zhao H, Wei T, Zhou W, Zhang W, Chen D, Yu N. Multi-attentional deepfake detection. In: Proc. of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: CVPR, 2021. 2185-2194.
[10] Chen L, Zhang Y, Song Y, Liu L, Wang J. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. arXiv preprint arXiv: 2203.12208. 2022.
[11] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations. In: Proc. of the 2019 IEEE Winter Applications of Computer Vision Workshops. Waikoloa: WACVW, 2019. 83-92.
[12] Korshunov P, Marcel S. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. arXiv preprint arXiv:
[13] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. Faceforensics++: Learning to detect manipulated facial images. In: Proc. of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: ICCV, 2019. 1-11.
[14] Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: A new dataset for Deepfake forensics. arXiv preprint arXiv:1909.12962. 2019.
[15] Jiang L, Li R, Wu W, Qian C, Loy CC. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In: Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: CVPR, 2020. 83-92.
[16] Zi BJ, Chang MH, Chen JJ, Ma XJ, Jiang YG. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. arXiv preprint arXiv: 2101.01456. 2021.
[17] Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang ML, Ferrer CC. The DeepFake Detection Challenge (DFDC) Dataset. arXiv preprint arXiv: 2006.07397. 2020.
[18] Patrick K, Jaeseong Y, Gyuhyeon N, Sungwoo P, Gyeongsu C. KoDF: A Large-scale Korean DeepFake Detection Dataset. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: ICCV, 2021. 10724-10733.
[19] He YA, Gan, B, Chen, SY, Zhou YC, Yin GJ, Song LCA, Sheng L, Shao J, Liu ZW. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In: Proc. of the 2021 IEEE Conf. on Computer Vision and Pattern Recognition. Nashville: CVPR.
[20] Khalid H, TariqS, Kim M, Simon S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. arXiv preprint arXiv:
[21] Chen C, Wang D, Zheng TF. CN-CVS: A Mandarin Audio-Visual Dataset for Large Vocabulary Continuous Visual to Speech Synthesis. In: Proc. of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes: ICASSP.
[22] Zhao Y, Xu R, ML Song. A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading. In: Proc. of the 1st ACM International Conference on Multimedia in Asia. New York: MMAsia '19. 2020. 1–6.

[23] Mockingbird. 2021. https://github.com/babysor/MockingBird
[24] Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N. First order motion model for image animation. In: Proc. of the Advances in Neural Information Processing Systems. Red Hook: NeurIPS. 2019. 7137–7147.
[25] Nirkin Y, Keller Y, Hassner T. Fsgan: Subject agnostic face swapping and reenactment. In: Proc. of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: ICCV. 2019. 7183-7192.
[26] Siarohin A, Roy S, Lathuilière S, Tulyakov S, Ricci E, Sebe N. Motion-supervised Co-Part Segmentation. In: Proc. of the 2020 25th International Conference on Pattern Recognition. Milan: ICPR. 2021. 9650-9657.
[27] Chen RW, Chen XH, Ni BB, Ge YH. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In Proc. of the 28th ACM International Conference on Multimedia. New York: MM '20. 2020. 2003–2011.
[28] Chung JS, Andrew Z. Out of time: Automated lip sync in the wild. In: Proc. of the Asian Conference on Computer Vision. 2017.
[29] coqui TTS. 2023. https://github.com/coqui-ai/TTS
[30] FakeApp. 2019. https://www.deepfakescn.com
[31] Faceswap: Deepfakes software for all. 2020. https://github.com/deepfakes/faceswap
[32] Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M. Face2face: Real-time face capture and reenactment of rgb videos. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. LasVegas: CVPR. 2016. 2387-2395.
[33] Faceswap. 2020. https://github.com/MarekKowalski/FaceSwap/
[34] Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-12.
[35] Petrov I, Gao DH, Chervoniy N, Liu K, Marangonda S, Chris U, Dpfks M, Luis RP, Jiang J, Zhang S, Wu PY, Zhou B, Zhang WM. Deepfacelab: A simple, flexible and extensible face swapping framework. arXiv preprint arXiv:2005.05535. 2020.
[36] Jia Y, Zhang Y, Weiss R, Wang Q, Shen J, Ren F, Chen ZF, Nguyen P, Pang RM, Moreno I, Wu YH. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Proc. of the 32nd International Conference on Neural Information Processing Systems. Red Hook: NIPS'18. 2018. 4485–4495.
[37] Wang YX, Skerry-Ryan R, Stanton D, Wu YH, Weiss R, Jaitly N, Yang ZH, Xiao Y, Chen ZF, Bengio S, Le Q, Agiomyrgiannakis Y, Clark R, Saurous R. Tacotron: Towards End-to-End Speech Synthesis. arXiv preprint arXiv: 1703.10135. 2017.
[38] Shen J, Pang R, Weiss R, Schuster M, Jaitly N, Yang ZH, Chen ZF, Zhang Y, Wang YX, Skerrv-Ryan R, Saurous R, Agiomvrgiannakis Y, Wu YH. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proc. of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: ICASSP. 2018. 4779–4783.
[39] Kim J, Kim S, Kong J, Yoon S. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search. In Proc. of the 34th International

Conference on Neural Information Processing Systems. Red Hook: NIPS'20. 2020. 8067–8077.
[40] Kumar K, Kumar R, Boissiere T, Gestin L, Teoh WZ, Sotelo J, Brebisson A, Bengio Y, Courville A. MelGAN: generative adversarial networks for conditional waveform synthesis. In the Proc. of the 33rd International Conference on Neural Information Processing Systems. Red Hook: NIPS. 2019. 14910–14921.
[41] Yang G, Yang S, Liu K, Fang P, Chen W, Xie L. Multi-Band Melgan: Faster Waveform Generation For High-Quality Text-To-Speech. In the Proc. of the 2021 IEEE Spoken Language Technology Workshop. Shenzhen: SLT. 2021. 492-498.
[42] Mikolaj B, Donahue J, Dieleman S, Clark A, Elsen E, Casagrande N, Luis CC, Karen S. High Fidelity Speech Synthesis with Adversarial Networks. arXiv preprint arXiv: 1909.11646. 2019.
[43] Zhang YB, Lin WG, and Xu JF. Joint Audio-Visual Attention with Contrastive Learning for More General Deepfake Detection. In the Proc. of the ACM Transactions on Multimedia Computing, Communications and Applications. TOMCCAP. 2023. Just Accepted
[44] Chao F, Chen ZY, Andrew O. Self-Supervised Video Forensics by Audio-Visual Anomaly Detection. In the Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: CVPR. 2023. 10491-10503.
[45] Liu XL, Yu Y, Li XL, Zhao Y. Magnifying multimodal forgery clues for Deepfake detection. Image Communication. 2023,118(C).
[46] Shahzad SA, Hashmi A, Khan S, Peng YT, Tsao Y, Wang HM. Lip Sync Matters: A Novel Multimodal Forgery Detector. In the Proc. of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Chiang Mai: APSIPA ASC. 2022. 1885-1892.
[47] Liu X, Yu Y, Li X, Zhao Y. MCL: Multimodal Contrastive Learning for Deepfake Detection. IEEE Transactions on Circuits and Systems for Video Technology. 2023. doi: 10.1109/TCSVT.2023.3312738.
[48] Cozzolino D, Pianese A, Nießner M, Verdoliva L. Audio-Visual Person-of-Interest DeepFake Detection. In the Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: CVPRW. 2023. 943-952.
[49] Yu Y, Liu X, Ni R, Yang S, Zhao Y, Kot AC. PVASS-MDD: Predictive Visual-audio Alignment Self-supervision for Multimodal Deepfake Detection. IEEE Transactions on Circuits and Systems for Video Technology. 2023. doi: 10.1109/TCSVT.2023.3309899.
[50] Ilyas H, Javed A, Malik KM. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. Applied Soft Computing. 136(C). 2023. doi:10.1016/j.asoc.2023.110124
[51] Hashmi A, S. Shahzad A, Ahmad W, Lin CW, Tsao Y, Wang HM. Multimodal Forgery Detection Using Ensemble Learning. In the Proc. of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Chiang Mai: APSIPA ASC. 2022. 1524-1532.
[52] Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: A compact facial

video forgery detection network. In Proc. of the 2018 IEEE International Workshop on Information Forensics and Security. Hong Kong: WIFS. 2018. 1-7.
[53] Christian S, Liu W, Jia YQ, Pierre S, Scott R, Dragomir A, Dumitru E, Vincent V, Andrew R. Going deeper with convolutions. In Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: CVPR. 2015. 1-9.
[54] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: CVPR. 2017. 1800-1807.
[55] Verma D 2021. https://github.com/dh1105/Multi-modal-movie-genre-prediction
[56] Yu ZT, Zhao CX, Wang ZZ, Qin YX, Su Z, Li XB, Zhou F, Zhao GY. Searching central difference convolutional networks for face anti-spoofing. In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: CVPR. 2020.
[57] Chugh K, Gupta P, Dhall A, Subramanian R. Not made for each other-audio-visual dissonance-based deepfake detection and localization. in Proc. of the 28th ACM International Conference on Multimedia. 2020. 439–447.
[58] Cheng H, Guo YY, Wang TY, Li Q, Chang XJ, Nie LQ. Voice-Face Homogeneity Tells Deepfake. ACM Transactions on Multimedia Computing, Communications, and Applications. 2023. 20(3):1-22.
[59] Yang WY, Zhou XY, Chen ZK, Guo BF, Ba ZJ, Xia ZH, Cao XC, Ren K. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. IEEE Transactions on Information Forensics and Security, 2023, 18:2015-2029.
[60] Khalid H, Kim M, S, Tariq, Woo S. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In Proc. of the 1st workshop on synthetic multimedia-audio visual deepfake generation and detection. New York: ADGD '21. 2021. 7–15.
[61] Xi S, JX Ma, Zhou C, Yang ZX. Controllable 3D Face Generation with Conditional Style Code Diffusion. arXiv preprint arXiv: 2312.13941, 2024.
[62] Qing ZW, Zhang SW, Wang JY, Wang X, Wei YJ, Zhang YY, Gao CX, Sang N. Hierarchical Spatio-temporal Decoupling for Text-to-Video Generation. arXiv preprint arXiv: 2312.04483, 2023.
[63] Zeng Y and Wei GQ, Zheng JN, Zou JX, Wei Y, Zhang YC, Li H. Make Pixels Dance: High-Dynamic Video Generation. arXiv preprint arXiv: 2311.10982, 2023.
[64] Ho J, Saharia C, Chan W, David J, Norouzi M, Salimans T. Cascaded Diffusion Models for High Fidelity Image Generation. arXiv preprint arXiv: 2106.15282, 2021.
[65] Li XR, Ji SL, Wu CM, Liu ZG, Deng SG, Cheng P, Yang M, Kong XW. Survey on Deepfakes and Detection Techniques. Ruan Jian Xue Bao/Journal of Software, 2021,32(2):496-518. (in Chinese with English abstract). http://www.jos.org.cn/jos/article/html/6140

Chinese Reference:

[65] Li Xurong, Ji Shouling, Wu Chunming, Liu Zhenguang, Deng Shuiguang, Cheng Peng, Yang Min, Kong Xiangwei. Survey on Deepfakes and Detection Techniques. Journal of Software, 2021,32(2):496-518. http://www.jos.org.cn/jos/article/html/6140

Note: Figure translations are in progress. See original paper for figures.