

Identity-Driven Multimedia Forgery Detection via Reference Assistance

Anonymous Authors

1 FORGERY DATA GENERATION

Our IDForge employ 7 types of forgery techniques:

- Lip synchronization (*lip*): This method aligns lip movements in a video to correspond with a new audio track. We employ Wav2Lip for precise lip synchronization.
- Face-swapping (*face*): This technique replaces an individual’s face in a video with another person’s. We implement face-swapping using advanced methods like InsightFace, SimSwap, and InfoSwap.
- Voice cloning (*tts*): Voice cloning generates new audio sequences from text, closely mimicking a person’s unique voice. In this process, we utilize TorToiSe for effective voice cloning.
- Voice conversion (*rvc*): This technique alters the voice of one individual to resemble another’s while keeping the original speech content intact. We achieve this through RVC.
- Audio shuffling (*audioshuffle*): In audio shuffling, we exchange the audio tracks between individuals of the same gender. This technique creates an effect similar to dubbed video content found online.
- LLM generation (*textgen*): This approach generates new text stylistically consistent with the original but conveys opposite or altered content. We leverage GPT-3.5 for this sophisticated text generation.
- Text shuffling (*textshuffle*): Text shuffling entails exchanging one individual’s transcript with another, producing fabricated yet human-originated texts.

Considering that real-world forgeries often involve multiple manipulations across different modalities, we combine forgery techniques from different modalities to obtain 11 distinct types of multimedia forgeries, as illustrated in Table 1. Among these, 4 multimedia forgeries are categorized as forgery-aligned forgeries. These represent the most deceptive forgeries on the internet, having undergone extensive manipulation in all three modalities. In contrast, we also have 7 non-forgery-aligned forgeries. These are commonly found online and involve partial manipulations. Although they may have fewer apparent defects, they still pose significant challenges for detection methods due to their subtler nature.

2 DATASET SPLIT

In the data generation stage, we select 6 wild videos from YouTube for each individual. For dataset splitting, video shots generated from four of these selected videos are allocated to the training set. In contrast, video shots generated from the remaining two are assigned to the test set. This split strategy mirrors real-world scenarios where data used for training media detection methods, and the data subjected to testing by these methods often originate from distinct domains. Consequently, videos sourced from different selected videos may exhibit varied characteristics, including background scenes, background noise, and topics. To split a small number of video shots for the validation set, we directly separate

Table 1: Combinations of different forgery techniques categories

Type	Forgery-aligned	Number
lip+tts+textgen	Yes	22,983
lip+rvc+textshuffle	Yes	22,985
face+tts+textgen	Yes	16,881
face+rvc+textshuffle	Yes	16,796
tts+textgen	No	15,985
tts+textshuffle	No	15,990
rvc+textshuffle	No	7,982
faceswap+audioshuffle+textshuffle	No	35,524
lip+audioshuffle+textshuffle	No	22,987
face+tts	No	9,936
pristine	-	79,827

video shots of 4 randomly chosen individuals as the validation set. Finally, we split IDForge into training, testing, and validation sets: 61.83% of video shots are in training sets, 6.95% for validation, and the remaining 31.22% for testing.

3 BASELINES

Our paper introduces several state-of-the-art methods on IDForge and FakeAVCeleb. The following are their detailed descriptions.

- **Xception** is an image-based method based on an Xception net that directly learns visual cues to determine whether a video is a forgery.
- **EfficientNet** is an image-based method based on an EfficientNet-B0 that directly leverages visual information for media forgery detection.
- **Meso-Inception4** is an image-based detection network which can efficiently detect face tampering in videos with a low computational cost.
- **P3D** is a video-based method based on the pseudo-3D residual network, which learns spatio-temporal video representation to detect forgeries.
- **I3D** is a video-based method based on the two-stream inflated 3D convolutional network, which extracts spatio-temporal features for media forgery detection.
- **CDCN** is an audio-visual method based on the central difference convolutional network, which is originally proposed for face anti-spoofing.
- **FTCN** is a video-based method comprised of a fully temporal convolution network and a temporal transformer network.
- **LVNet** is an image-based approach designed to enhance the robustness and generalizability of deepfake detection by explicitly identifying and focusing on potential forged regions within images.

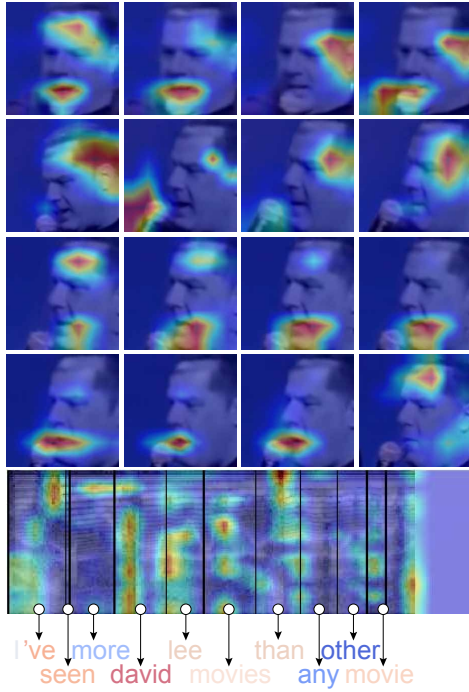


Figure 1: Class activation maps (CAMs) of R-MFDN on the IDForge dataset are displayed. The top four rows display the CAMs of grouped frames; the middle row shows the CAM of the audio spectrogram; and the bottom row presents the CAMs for the transcript tokens. The audio spectrogram is separated along the time dimension according to the word timestamps of the transcript tokens.

- **UCF** is an image-based approach designed to enhance the generalizability of deepfake detection by decomposing image information into distinct components: forgery-irrelevant, method-specific forgery, and common forgery features.
- **RawNet2-ASVspoof** is a network modified from original RawNet2 architecture to detect synthetic speech in ASVspoof challenge.
- **VFD** is an audio-visual method which is pre-trained and fine-tuned using both generic and deepfake audio-visual datasets. It detects deepfakes by extracting face images and voice clips and then assesses their match using a joint latent space. The

video is classified as real with matched voice-face pairs if their similarity surpasses a certain threshold.

- **Joint AV** is also an audio-visual method using a sync stream that models the synchronization patterns of two modalities to detect the forgeries.
- **ICT-Ref** is a method using an Identity Consistency Transformer to learn the identities in the inner face and outer face. The method detects the forgeries by measuring the inconsistency between the two identities.
- **ID-Reveal** contains a 3D Morphable Model (3DMM) for frame representation and a Temporal ID Network for embedding vectors, enhanced by a 3DMM Generative Network trained adversarially to include behavioral data. It authenticates videos by comparing these vectors with reference videos, assessing authenticity through behavioral and visual similarities.
- **POI-Forensics** is an audio-visual method trained on real videos to learn a specific identity representation to reveal if the identity of the video under test is real. The training stage employs a contrastive learning paradigm, training on real talking-face videos to extract visual and audio features that characterize the identity of a given person.
- **RealForensics** is an audio-visual method that consists of two stages. In the first stage, the model is trained on online videos of real faces for self-supervised learning of video representations. In the second stage, the model integrates visual and auditory features and is fine-tuned for binary real/fake classification.

It is worth noting that image-based methods treat each image as an individual classification task. Therefore, in our implementation, as we extract four groups of 4 successive frames for the video-based model, we calculate the average of the logits obtained from these image-based methods for these 16 frames to represent the final result of the corresponding video.

4 VISUALIZATION

Figure 1 shows the class activation maps of R-MFDN for the frames, audio spectrogram, and transcript tokens, respectively. The heatmap indicates that our method primarily focuses on the lip region and the outer edge of the face at the visual level. In R-MFDN, we introduce a cross-modal contrastive learning strategy to capture the inconsistencies across modalities. In Figure 1, the spectrogram and transcripts share a similarly focused region along the time dimension, indicating that the audio and textual encoders successfully align through cross-modal contrastive learning.