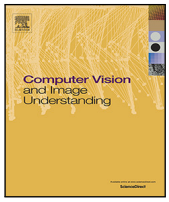
Computer Vision and Image Understanding 236 (2023) 103818

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization

Zhixi Cai [a,*], Shreya Ghosh [b], Abhinav Dhall [c], Tom Gedeon [b], Kalin Stefanov [a], Munawar Hayat [a]

[a] *Monash University, Clayton, Australia*
[b] *Curtin University, Perth, Australia*
[c] *Indian Institute of Technology Ropar, Ropar, India*

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Datasets
Deepfake
Localization
Detection

ABSTRACT

Most deepfake detection methods focus on detecting spatial and/or spatio-temporal changes in facial attributes and are centered around the binary classification task of detecting whether a video is real or fake. This is because available benchmark datasets contain mostly visual-only modifications present in the entirety of the video. However, a sophisticated deepfake may include small segments of audio or audio–visual manipulations that can completely change the meaning of the video content. To addresses this gap, we propose and benchmark a new dataset, Localized Audio Visual DeepFake (LAV-DF), consisting of strategic content-driven audio, visual and audio–visual manipulations. The proposed baseline method, Boundary Aware Temporal Forgery Detection (BA-TFD), is a 3D Convolutional Neural Network-based architecture which effectively captures multimodal manipulations. We further improve (i.e. BA-TFD+) the baseline method by replacing the backbone with a Multiscale Vision Transformer and guide the training process with contrastive, frame classification, boundary matching and multimodal boundary matching loss functions. The quantitative analysis demonstrates the superiority of BA-TFD+ on temporal forgery localization and deepfake detection tasks using several benchmark datasets including our newly proposed dataset. The dataset, models and code are available at https://github.com/ControlNet/LAV-DF.

1. Introduction

Increasingly powerful deep learning algorithms (e.g. Autoencoders Rumelhart et al., 1985 and Generative Adversarial Networks Goodfellow et al., 2020) accompanied by the rapid advances in computing power have enabled the generation of highly realistic synthetic media commonly referred to as *deepfakes*.[1] Audio–visual deepfake content generation utilizes methods for voice cloning (Wang et al., 2017; Jia et al., 2018), face reenactment (Tulyakov et al., 2018; Prajwal et al., 2020), and face swapping (Korshunova et al., 2017; Nirkin et al., 2019).

Audio–visual deepfakes include videos that have been either manipulated or created from scratch to primarily mislead, deceive or influence audiences. Given that access to deepfake generation technologies has become widespread and the technologies are easy to use, some researchers argue that deepfakes are "threat to democracy" (Schwartz, 2018; Brandon, 2019; Sample, 2020; Thomas, 2020). For example, Thies et al. (2020) used a video of the former United States president Barack Obama to demonstrate a novel face reenactment method. In the resultant realistic video, the former president's lip movement is synchronized with the speech of another person. This type of manipulations has the potential to mislead people in forming wrong opinions and could have serious consequences.

Given the rapid grow of fake videos on the Internet, robust and accurate deepfake detection methods are increasingly important. This triggered the release of several benchmark datasets for deepfake detection (Korshunov and Marcel, 2018; Rossler et al., 2019; Dolhansky et al., 2020; He et al., 2021) and state-of-the-art deepfake detection methods (Chen et al., 2022; Raza and Malik, 2023; Ilyas et al., 2023; Bayar and Stamm, 2016; Cozzolino et al., 2017; Yang et al., 2019; Li et al., 2020a) demonstrate promising performance on those benchmark datasets, which define the problem as a binary classification task (i.e. classify the whole input video as *real* or *fake*).

Fake content however, may only constitute a small part(s) of a long real video (Chugh et al., 2020) and these modified segment(s) could completely change the meaning and sentiment of the original content. Lets consider the example illustrated in Fig. 1, where the real video on the left captures the person saying "Vaccinations are safe". When the word "safe" is replaced with its antonym "dangerous", the meaning and sentiment of the video is significantly changed. This type of video forgeries can effectively manipulate the public opinion, particularly

* Corresponding author.
E-mail address: zhixi.cai@monash.edu (Z. Cai).

[1] In the text, *deepfake* and *forgery* are used interchangeably.

https://doi.org/10.1016/j.cviu.2023.103818
Received 5 May 2023; Received in revised form 16 July 2023; Accepted 21 August 2023
Available online 23 August 2023

Table 1
Details for publicly available deepfake datasets in a chronologically ascending order. The LAV-DF dataset details are reported in the last row. *Cla: Classification, SL: Spatial Localization, TFL: Temporal Forgery Localization, FS: Face Swapping,* and *RE: ReEnactment.*

Dataset	Year	Tasks	Manipulated modality	Manipulation method	#Subjects	#Real	#Fake	#Total
DF-TIMIT (Korshunov and Marcel, 2018)	2018	Cla	V	FS	43	320	640	960
UADFV (Yang et al., 2019)	2019	Cla	V	FS	49	49	49	98
FaceForensics++ (Rossler et al., 2019)	2019	Cla	V	FS/RE	–	1000	4000	5000
Google DFD (Nick and Andrew, 2019)	2019	Cla	V	FS	–	363	3068	3431
DFDC (Dolhansky et al., 2020)	2020	Cla	AV	FS	960	23,654	104,500	128,154
DeeperForensics (Jiang et al., 2020)	2020	Cla	V	FS	100	50,000	10,000	60,000
Celeb-DF (Li et al., 2020b)	2020	Cla	V	FS	59	590	5639	6229
WildDeepfake (Zi et al., 2020)	2020	Cla	–	–	–	3805	3509	7314
$FFIW_{10K}$ (Zhou et al., 2021b)	2021	Cla	V	FS	–	10,000	10,000	20,000
KoDF (Kwon et al., 2021)	2021	Cla	V	FS/RE	403	62,166	175,776	237,942
FakeAVCeleb (Khalid et al., 2021b)	2021	Cla	AV	RE	600+	570	25,000+	25,500+
ForgeryNet (He et al., 2021)	2021	SL/TFL/Cla	V	Random FS/RE	5400+	99,630	121,617	221,247
DF-Platter (Narayan et al., 2023)	2023	Cla	V	FS	454	133,260	132,496	265,756
LAV-DF (ours)	2022	TFL/Cla	AV	Content-driven RE	153	36,431	99,873	136,304

when targeting media involving famous individuals, as the example with Barack Obama. Given the underlying assumption (i.e. deepfake detection is a binary classification problem) of the current deepfake detection benchmark datasets and methods, it is possible that the state-of-the-art techniques may not perform well in identifying this new type of manipulations.

This paper addresses the important task of content-driven forgery localization and detection in video. In terms of benchmark datasets, there is a significant gap in the availability of datasets for multimodal content-driven forgery localization and detection. This paper proposes a pipeline for generating such large-scale dataset that can serve as a valuable resource for future research in this area. Furthermore, this paper also introduces a novel multimodal method that utilizes audio and visual information to precisely detect the boundaries of fake segments in videos. The **main contributions** of our work are,

- A large-scale public dataset, *Localized Audio Visual DeepFake*, for temporal forgery localization and detection.
- A multimodal method, *Boundary Aware Temporal Forgery Detection+*, for fake segment localization and detection.
- A thorough validation of the method's components and comprehensive comparison with the state-of-the-art.

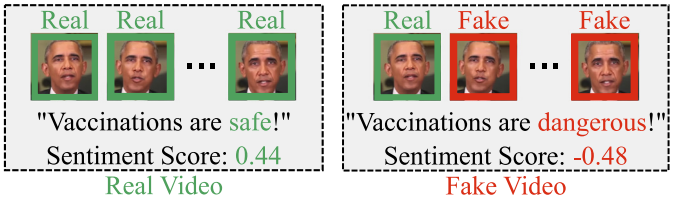


Fig. 1. Content-driven audio–visual manipulation. In the real video (left) the subject is saying "Vaccinations are safe". In the audio–visual deepfake (right) created from the real video, "safe" is changed to "dangerous" (resulting in a significant change in perceived sentiment). The green-edge frames are real and red-edge are fake. *Note that through a subtle audio–visual manipulation, the meaning of the video content has been completely changed*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Related work

This section reviews the relevant literature on deepfake detection datasets and methods. Given the similarities between temporal forgery localization and temporal action localization, previous work in the latter area is also reviewed.

2.1. Deepfake detection datasets

Deepfake detection research is driven by datasets generated with various deepfake generation approaches. We present a summary of the deepfake detection datasets available to the research community in Table 1. The first deepfake dataset named DF-TIMIT was proposed by Korshunov and Marcel (2018). DF-TIMIT curation process involved face swapping on VidTimit dataset (Sanderson, 2002). Later, UADFV (Yang et al., 2018), FaceForensics++ (Rossler et al., 2019) and Google DFD (Nick and Andrew, 2019) were introduced, and FaceForensics++ has become a popular benchmark dataset for multiple deepfake detection methods (Wang et al., 2022; Qian et al., 2020). The main limitation of the aforementioned datasets is their size (i.e. a maximum of thousands of video samples). Given that CNNs and Transformers (commonly used for deeepfake detection) are data-demanding techniques, these datasets have low generalization capability (Li et al., 2020b). In 2020, Facebook (i.e. Meta) published the large-scale dataset DFDC (Dolhansky et al., 2020) for deepfake detection with more than 100 K samples. Until today, DFDC is the standard benchmark used for deepfake detection methods (Yang et al., 2023b; Chen et al., 2022). After DFDC, several datasets targeting different specializations were introduced. For example, WildDeepfake (Zi et al., 2020) for web-crawled in-the-wild fake video detection, $FFIW_{10K}$ (Zhou et al., 2021b) for detecting fake faces in videos containing multiple faces, KoDF (Kwon et al., 2021) for Korean deepfake detection, and DF-Platter (Narayan et al., 2023) for detecting multi-face heterogeneous deepfakes. DeeperForensics (Jiang et al., 2020) is another notable dataset that overcomes the bias of having high number of fake videos. However, all those datasets mainly consider visual-only deepfake detection. In 2021, FakeAVCeleb (Khalid et al., 2021b) was introduced including both face swapping and audio-based face reenactment. This dataset includes fake audio generated from SV2TTS (Jia et al., 2018), which makes it the first deepfake detection dataset focusing on audio–visual manipulations.

Given that all of the those datasets regard the deepfake detection as a binary classification problem, ForgeryNet (He et al., 2021) dataset was introduced, which includes visual-only face swapping in random frames and is suitable for both video/image classification and spatial/temporal forgery localization. However, ForgeryNet only applies random face swapping in the visual modality and does not consider audio and content-driven modifications. To bridge this gap, we propose a multimodal content-driven temporal forgery localization and detection dataset.

2.2. Deepfake detection methods

Deepfake detection methods can be categorized into two categories: traditional machine learning and deep learning approaches. The traditional machine learning methods include EM (Guarnera et al., 2020)

and SVM (Yang et al., 2019). On the other hand, deep learning methods include CNN (de Lima et al., 2020), RNN (Montserrat et al., 2020; Chen et al., 2022) and ViT (Wodajo and Atnafu, 2021; Heo et al., 2023; Coccomini et al., 2022). Most of the prior deepfake detection methods focus on temporal inconsistencies (Lewis et al., 2020; Gu et al., 2021) and multimodal synchronization (Chugh et al., 2020; Wang et al., 2022; Mittal et al., 2020; Zhu et al., 2023) to detect fake videos.

All of the above mentioned methods employ classification centric approach. Thus, those methods do not have temporal localization capabilities. Only MDS (Chugh et al., 2020) demonstrated scenarios where only parts of the video are modified, although this approach is primarily designed for classification. Our dataset and method are designed to consider both audio–visual deepfake detection and temporal localization.

2.3. Temporal action localization

Since the temporal forgery localization task is similar to temporal action localization, we also review the literature in this domain. For temporal action localization, ActivityNet (Caba Heilbron et al., 2015), THUMOS14 (Idrees et al., 2017), HACS (Zhao et al., 2019), EPIC-KITCHEN (Damen et al., 2022), and FineAction (Liu et al., 2022c) are popular benchmark datasets. Temporal action localization methods can be classified as two types: 2-stage approaches (Zeng et al., 2019; Xu et al., 2020; Liu et al., 2021), where the temporal bounding box proposals are generated at first and then are classified as different classes, and 1-stage approaches (Lin et al., 2017; Buch et al., 2019; Nawhal and Mori, 2021; Zhang et al., 2022; Liu et al., 2022a,b; Shi et al., 2023; Yang et al., 2023a), which directly predict the final temporal segments.

For temporal forgery localization, there is no requirement to classify the foreground segments, in other words, the background is always real and the foreground is always fake. Hence, 1-stage temporal action localization approaches are more relevant for the task. According to Bagchi et al. (2022), these approaches can be grouped in two main categories: methods based on anchors and methods based on predicting the boundary probabilities. Anchor-based methods (Shou et al., 2016, 2017; Gao et al., 2017, 2018) utilize sliding windows in the video to detect segments. Lin et al. (2018) proposed a new framework to generate proposals that predicts the boundary probabilities based on start and end timestamps. This approach can access the global context information to generate more precise and flexible segment proposals than anchor-based methods. Based on this method, several other approaches were proposed to enhance performance (Lin et al., 2019; Su et al., 2021).

All temporal action localization methods described above are visual-only, which is not optimal for the task of temporal forgery localization. The importance of accessing the multimodal information for temporal action localization was recently raised by Bagchi et al. (2022).

2.4. Proposed multimodal approach

This paper proposes a multimodal method for precise boundary proposal estimation to detect and localize fake segments videos. We quantitatively compare the performance of the proposed method with existing state-of-the-art approaches, including BMN (Lin et al., 2019), AGT (Nawhal and Mori, 2021), MDS (Chugh et al., 2020), AVFusion (Bagchi et al., 2022), BSN++ (Su et al., 2021), TadTR (Liu et al., 2022b), ActionFormer (Zhang et al., 2022), and TriDet (Shi et al., 2023).

3. Localized audio visual DeepFake dataset

We created a large-scale audio–visual deepfake dataset containing 136,304 videos (36,431 real and 99,873 fake). Our data generation pipeline is illustrated in Fig. 2. The generation is guided by relevant words in the video transcripts and specifically, the manipulation strategy is to replace strategic words with their antonyms, which leads to a significant change in the perceived sentiment of the statement.

3.1. Audio–visual data sourcing

The real videos in this dataset are collected from the VoxCeleb2 dataset (Chung et al., 2018), which is a large-scale facial video dataset containing more than 1 million utterance videos of 6112 speakers. To ensure consistency, the faces within these videos are tracked and cropped using the Dlib facial detector (King, 2009) at 224 × 224 resolution. The VoxCeleb2 dataset offers a diverse range of video lengths, spoken languages, and voice qualities. Our dataset includes only English-speaking videos, where the spoken language was detected through the confidence score generated by the Google Speech-to-Text service.[2] We leveraged the same service to generate the transcripts.

3.2. Audio–visual data generation

After sourcing the real videos, the next step is to analyze each video transcript for content-driven deepfake generation. The generation process includes transcript manipulation, followed by generation of the corresponding audio and visual modalities.

3.2.1. Transcript manipulation

Following the collection and wrangling of the real data, the next step is to analyze the transcript of a video denoted as $D = \{d_0, d_1, \ldots, d_m, \ldots, d_n\}$, where d_i represents individual word tokens and n denotes the total number of tokens in the transcript. The objective is to identify the tokens within D that should be replaced in order to achieve the maximum change in perceived sentiment. This process aims to create a modified transcript $D' = \{d_0, d_1, \ldots, d'_m, \ldots, d_n\}$, which consists of most of the original tokens from D and the replacements for a few specific tokens. The replacement tokens, denoted as d', are selected from a set $\hat{d}$ containing antonyms of d, sourced from WordNet (Fellbaum, 1998). To determine the sentiment value of the transcript, we employed the sentiment analyzer available in NLTK (Bird et al., 2009). Specifically, for each token d in a transcript D, the replacement is found with,

$$\tau = \underset{d \in D, d' \in \hat{d}}{\operatorname{argmax}} |S(D) - S(D')|$$

Then all replacements in a transcript D are found as follows,

$$\theta = \underset{\{\tau_m\}_{m=1}^{M}}{\operatorname{argmax}} \left| \sum_{i=1}^{M} \Delta S(\tau_i) \right|$$

where $\Delta S(\tau_i)$ is the difference in sentiment score of the original and modified transcripts when utilizing the replacement τ_i and M is the maximum number of replacements.

There is up to 1 replacement for videos shorter than 10 s; otherwise, there can be a maximum of 2 replacements. The shift in sentiment distribution following the manipulations is visualized in Fig. 3(a), while the histogram of $|\Delta S|$ indicating that the sentiment of most transcripts has been successfully changed, is shown in Fig. 3(b).

3.2.2. Audio generation

After the transcript manipulation, the next step is to generate speaker-specific audio for the replacement tokens. Motivated by the prior work on adaptive text-to-speech methods (Jia et al., 2018; Casanova et al., 2021; Neekhara et al., 2021), we adopted SV2TTS (Jia et al., 2018) for speaker-specific audio generation. SV2TTS consists of three modules: (1) An encoder module responsible for extracting the style embedding of the reference speaker, (2) A spectrogram generation module based on Tacotron 2 (Shen et al., 2018) utilizing replacement tokens and the speaker style embedding, and (3) A vocoder module based on WaveNet (Oord et al., 2016), which generates realistic audio using the spectrogram. In the audio generation, we utilized a pre-trained SV2TTS model to generate the audio segments. Then, we

[2] https://cloud.google.com/speech-to-text.

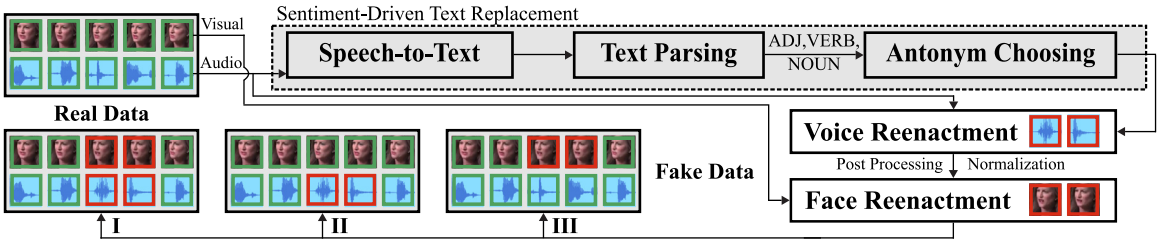


Fig. 2. Content-driven audio–visual manipulation for the creation of the LAV-DF dataset. The real transcript is used to find the word tokens for replacement based on the largest change in perceived sentiment. Then the modified tokens are used as input for generating audio. Post-processing and normalization are applied to the generated audio to maintain loudness consistency in the temporal neighborhood. The generated audio is then used as input for facial reenactment. The green-edge audio and visual frames are real data, and red-edge are fake data. *In total, three categories of data are generated: Fake Audio and Fake Visual, Fake Audio and Real Visual and Real Audio and Fake Visual*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

performed loudness normalization on the generated audio segments by considering the corresponding real audio neighbors. The rationale behind the loudness normalization is to generate a more realistic counterpart of the audio segment chosen for replacement.

3.2.3. Video generation

The generated audio is used as input for generating the corresponding visual frames. Wav2Lip (Prajwal et al., 2020) facial reenactment is used for this task, as it has been shown to achieve state-of-the-art output generation quality along with better generalization (Jamaludin et al., 2019; K R et al., 2019). We encountered several issues with using other popular visual generation methods such as AD-NeRF (Guo et al., 2021) and ATVGnet (Chen et al., 2019). For example, AD-NeRF does not fit in our generation context (i.e. zero-shot generation of unseen speakers), and ATVGnet uses a static reference image as input for facial reenactment, resulting in pose inconsistencies on the boundaries between real and fake segments. In contrast, Wav2Lip uses a reference video and target audio as input and generates an output video in which the person in the reference video lip-syncs to the target audio content, ensuring pose consistency between real and fake segments. We employed a pre-trained Wav2Lip model and upscaled the generated visual segments to a resolution of 224 × 224. The generated audio–visual segments are then synchronized and used to replace the original audio–visual segments.

Similar to Khalid et al. (2021a), LAV-DF includes three categories of generated data,

- **Fake Audio** and **Fake Visual.** Both the real audio and visual segments corresponding to the replacement tokens are manipulated.
- **Fake Audio** and **Real Visual.** Only the real audio segments corresponding to the replacement tokens are manipulated. To keep the fake audio and real visual segments synchronized, the corresponding real visual segments are length-normalized.
- **Real Audio** and **Fake Visual.** Only the real visual segments corresponding to the replacement tokens are manipulated and the length of the fake visual segments is normalized to match the length of the real audio segments.

3.3. Dataset statistics

The dataset contains 136,304 videos of 153 unique identities, with 36,431 real videos and 99,873 videos containing fake segments. For benchmarking, we splitted the dataset into 3 identity-independent subsets: train (78,703 videos of 91 identities), validation (31,501 videos of 31 identities), and test (26,100 videos of 31 identities). Summary of main statistics of the dataset is presented in Fig. 3.

The dataset includes a total of 114,253 fake segments, with duration $(0, 1.6]$ s and an average length of 0.65 s. Notably, 89.26% of the fake segments are shorter than 1 s. The maximum length of the videos in the dataset is 20 s and 69.61% of the videos are shorter than 10 s. In terms of modality modification, the distribution is balanced among the four types: visual-modified, audio-modified, both-modified and real. Additionally, the majority of the videos (62.72%) contain only 1 fake segment, while a smaller proportion of videos (10.55%) include 2 fake segments.

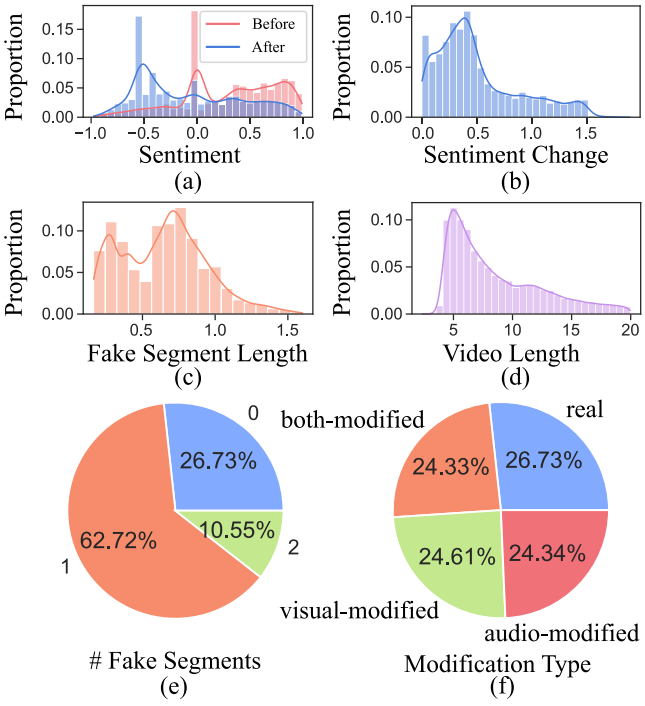


Fig. 3. Statistics of the LAV-DF dataset. (a) Distribution of sentiment scores before and after content-driven deepfake generation, (b) Histogram of sentiment changes $|\Delta S|$, (c) Distribution of fake segment lengths, (d) Distribution of video lengths, (e) Proportion of number of fake segments, and (f) Proportion of modifications.

3.4. Dataset quality

Table 2 provides a quantitative comparison (PSNR and SSIM) with existing dataset generation pipelines in terms of visual quality, demonstrating that our pipeline achieves better visual quality on the VoxCeleb2 dataset.

4. Boundary aware temporal forgery detection+ method

The objective is to detect and localize multimodal manipulations given an input video. To this end, we designed the proposed method BA-TFD+ in such a way that it has the capability to capture deepfake artifacts and localize the boundary of fake segments. An overview of the proposed method is depicted in Fig. 4 and Algorithm 1.

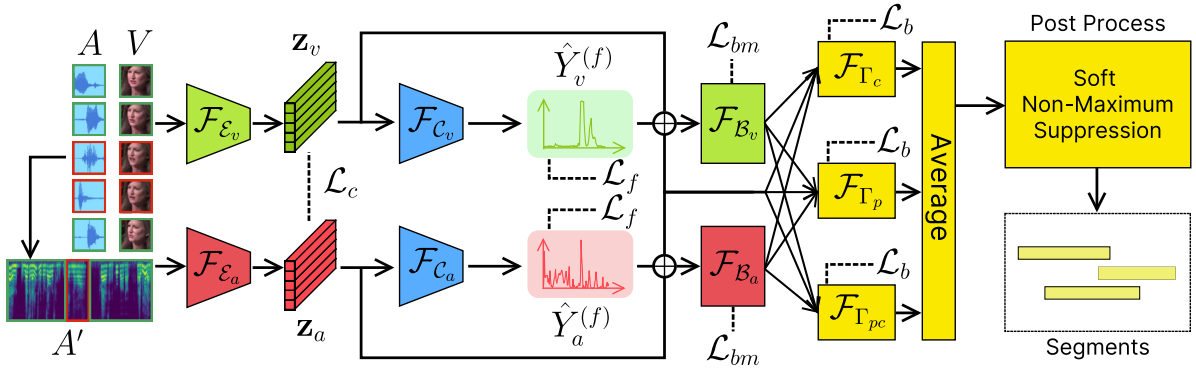


Fig. 4. Overview of the BA-TFD+ method. BA-TFD+ mainly comprises of (1) Visual encoder ($\mathcal{F}_{\mathcal{E}_v}$) that takes resized raw video frames as input, (2) Audio encoder ($\mathcal{F}_{\mathcal{E}_a}$) that takes spectrogram extracted from raw audio as input, (3) Visual and audio based frame classification module (i.e. $\mathcal{F}_{\mathcal{C}_v}$ and $\mathcal{F}_{\mathcal{C}_a}$), (4) Boundary localization module to facilitate forgery localization in both visual ($\mathcal{F}_{\mathcal{B}_v}$) and audio ($\mathcal{F}_{\mathcal{B}_a}$) modality, and finally (5) Multimodal fusion module that fuses multimodal latent features position-wise (p), channel-wise (c) and position-channel wise (pc). During inference, post-processing operation is applied to generate segments from the output of the fusion module. $\oplus$ *denotes concatenation.*

Algorithm 1: Training procedure of BA-TFD+

Data: Training data $\mathbb{D} \supset \{X_i, Y_i\}_{i=1}^{n}$, Modality modification flag $\mathbb{E} = \{\eta_i = (\eta_{vi}, \eta_{ai})\}_{i=1}^{n}$, Weights of losses λ
Result: Parameters of the model θ
$\theta \leftarrow$ Initialize the parameters randomly;
$Y_0 \leftarrow$ label for real data;
while θ *not converged* **do**
  $(V, A, Y) \leftarrow$ Next sample from $\mathbb{D}$;
  $(\eta_v, \eta_a) \leftarrow$ Next flag from $\mathbb{E}$;
  $Y_v \leftarrow$ if η_v then Y else Y_0;
  $Y_a \leftarrow$ if η_a then Y else Y_0;
  $Y^{(b)} \leftarrow$ Generate labels from Y;
  $(Y_v^{(b)}, Y_v^{(f)}) \leftarrow$ Generate labels from Y_v;
  $(Y_a^{(b)}, Y_a^{(f)}) \leftarrow$ Generate labels from Y_a;
  $\mathbf{z}_v \leftarrow \mathcal{F}_{\mathcal{E}_v}(V)$;
  $\mathbf{z}_a \leftarrow \mathcal{F}_{\mathcal{E}_a}(\text{mel-spectrogram}(A))$;
  $Y^{(c)} \leftarrow \eta_{vi} \wedge \eta_{ai}$;
  $\mathcal{L}_c \leftarrow$ ContrastiveLoss$(\mathbf{z}_v, \mathbf{z}_a, Y^{(c)})$;
  $\hat{Y}_v^{(f)} \leftarrow \mathcal{F}_{\mathcal{C}_v}(\mathbf{z}_v)$;
  $\hat{Y}_a^{(f)} \leftarrow \mathcal{F}_{\mathcal{C}_a}(\mathbf{z}_a)$;
  /* FL: Frame Loss */
  $\mathcal{L}_f \leftarrow \frac{1}{2}(\text{FL}(\hat{Y}_v^{(f)}, Y_v^{(f)}) + \text{FL}(\hat{Y}_a^{(f)}, Y_a^{(f)}))$;
  /* $\oplus$: concatenation */
  $(\hat{Y}_v^{(b)(p)}, \hat{Y}_v^{(b)(c)}, \hat{Y}_v^{(b)(pc)}) \leftarrow \mathcal{F}_{\mathcal{B}_v}(\mathbf{z}_v \oplus \hat{Y}_v^{(f)})$;
  $(\hat{Y}_a^{(b)(p)}, \hat{Y}_a^{(b)(c)}, \hat{Y}_a^{(b)(pc)}) \leftarrow \mathcal{F}_{\mathcal{B}_a}(\mathbf{z}_a \oplus \hat{Y}_a^{(f)})$;
  $\hat{Y}^{(b)(p)} \leftarrow \mathcal{F}_{\Gamma_p}(\hat{Y}_v^{(b)(p)}, \hat{Y}_a^{(b)(p)}, \mathbf{z}_v, \mathbf{z}_a)$;
  $\hat{Y}^{(b)(c)} \leftarrow \mathcal{F}_{\Gamma_c}(\hat{Y}_v^{(b)(c)}, \hat{Y}_a^{(b)(c)}, \mathbf{z}_v, \mathbf{z}_a)$;
  $\hat{Y}^{(b)(pc)} \leftarrow \mathcal{F}_{\Gamma_{pc}}(\hat{Y}_v^{(b)(pc)}, \hat{Y}_a^{(b)(pc)}, \mathbf{z}_v, \mathbf{z}_a)$;
  $\mathcal{L}_{bm} \leftarrow \frac{1}{2}(MSE(\hat{Y}_v^{(b)(p)}, Y_v^{(b)}) + MSE(\hat{Y}_v^{(b)(c)}, Y_v^{(b)}) + MSE(\hat{Y}_v^{(b)(pc)}, Y_v^{(b)}) + MSE(\hat{Y}_a^{(b)(p)}, Y_a^{(b)}) + MSE(\hat{Y}_a^{(b)(c)}, Y_a^{(b)}) + MSE(\hat{Y}_a^{(b)(pc)}, Y_a^{(b)}))$;
  $\mathcal{L}_b \leftarrow MSE(\hat{Y}^{(b)(p)}, Y^{(b)}) + MSE(\hat{Y}^{(b)(c)}, Y^{(b)}) + MSE(\hat{Y}^{(b)(pc)}, Y^{(b)})$;
  $\theta \leftarrow Adam(\mathcal{L}_b, \mathcal{L}_{bm}, \mathcal{L}_f, \mathcal{L}_c, \lambda, \theta)$;
end
return θ;

4.1. Preliminaries

The training dataset $\mathbb{D} \supset \{X_i, Y_i\}_{i=1}^{n}$ comprises of n multimodal inputs X_i with visual modality V_i and audio modality A_i, and the associated output labels Y_i. The proposed model BA-TFD+ with trainable parameters θ is optimized to map the inputs X_i to the outputs Y_i. Each X_i has a different number of frames t_i. In order to simplify the batch training of the model, we padded the temporal axis for all X_i to T.

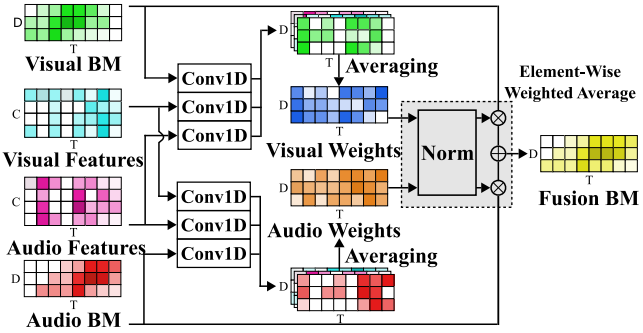


Fig. 5. Overview of the BA-TFD+ fusion module. The gray block is used to normalize the visual and audio weights produced by the 1D convolutional layers, followed by element-wise weighted average. $\oplus$ *denotes element-wise addition,* $\otimes$ *denotes element-wise multiplication and BM denotes boundary map.*

Table 2
Visual quality of the LAV-DF dataset. We maintained the experimental protocol and adopted the scores on VoxCeleb2 for the related deepfake generation pipelines from Zhou et al. (2021a).

Method	PSNR	SSIM
ATVGnet (Chen et al., 2019)	29.41	0.826
Wav2Lip (Prajwal et al., 2020)	29.54	0.846
MakeitTalk (Zhou et al., 2020)	29.51	0.817
Rhythmic Head (Chen et al., 2020)	29.55	0.779
PC-AVS (Zhou et al., 2021a)	29.68	0.886
LAV-DF (Ours)	**33.06**	**0.898**

4.2. Visual encoder

The goal of the visual encoder $\mathcal{F}_{\mathcal{E}_v}$ is to capture the frame-level spatio-temporal features from the input visual modality $V \supset \{V_i\}_{i=1}^{n}$ using an MViTv2 (Li et al., 2022). MViTv2 achieves seminal performance gain for different video analysis tasks including video action recognition and detection. In addition, MViTv2 leverages hierarchical multi-scale features compared to the basic ViT (Dosovitskiy et al., 2021). Our backbone MViTv2-Base model comprises of 4 blocks and 24 multi-head self-attention layers. As illustrated in Fig. 4, the visual encoder $\mathcal{F}_{\mathcal{E}_v}$ maps the inputs $V \in \mathbb{R}^{C \times T \times H \times W}$ (T is the number of frames, C is the number of channels, and H and W are the height and width of the frames) to latent space $\mathbf{z}_v \in \mathbb{R}^{C_f \times T}$ (C_f is the dimension of the features).

4.3. Audio encoder

The goal of the ViT-based (Dosovitskiy et al., 2021) audio encoder $\mathcal{F}_{\mathcal{E}_a}$ is to learn meaningful features from the raw input audio modality $A \supset \{A_i\}_{i=1}^{n}$. Following previous work (Ilyas et al., 2023; Yang et al., 2023b), we pre-process the raw audio A to generate representative mel-spectrograms $A' \in \mathbb{R}^{F_m \times T_a}$ ($T_a = \tau T$ is the temporal dimension and $\tau \in \mathbb{N}^*$, $\mathbb{N}^*$ denotes positive integers, and F_m is the length of the mel-frequency cepstrum features). In order to keep the audio–visual synchronization, we reshape the temporal axis of the mel-spectrograms to $\tau F_m \times T$. The reshaped spectrograms A' are given as input to the ViT blocks of the audio encoder $\mathcal{F}_{\mathcal{E}_a}$.[3] The audio encoder $\mathcal{F}_{\mathcal{E}_a}$ maps the mel-spectrograms A' to the latent space $\mathbf{z}_a \in \mathbb{R}^{C_f \times T}$, where C_f is the features dimension.

4.4. Frame classification module

We further deploy frame-level classification modules on top of the visual and audio features. Let us denote the ground truth labels for visual and audio modality as $Y_v^{(f)}$ and $Y_a^{(f)}$. The visual classification module $\mathcal{F}_{\mathcal{C}_v}$ maps the latent visual features $\mathbf{z}_v$ to labels $\hat{Y}_v^{(f)} \in \mathbb{R}^T$. Similarly, the audio classification module $\mathcal{F}_{\mathcal{C}_a}$ maps latent audio features $\mathbf{z}_a$ to labels $\hat{Y}_a^{(f)} \in \mathbb{R}^T$.

4.5. Boundary localization module

This module facilitates the learning of deepfake localization. Motivated by BSN++ (Su et al., 2021), we adopted the proposal relation block (PRB) as the framework for the boundary maps (representation of the boundary information of all densely distributed proposals). The ground truth boundary map $Y^{(b)} \in \mathbb{R}^{D \times T}$ is generated from Y, where $Y_{ij}^{(b)}$ is the confidence score for a segment which starts at the jth frame and ends at the $(i+j)$-th frame. The PRB module contains both a position-aware attention module (captures global dependencies) and a channel-aware attention module (captures inter-dependencies between different channels). In order to achieve localization in each modality, we deploy two boundary modules, $\mathcal{F}_{\mathcal{B}_v}$ for visual and $\mathcal{F}_{\mathcal{B}_a}$ for audio modality.

The visual boundary module $\mathcal{F}_{\mathcal{B}_v}$ input consists of the concatenation of latent features $\mathbf{z}_v$ and classification outputs $\hat{Y}_v^{(f)}$, i.e $\mathbf{z}_v \oplus \hat{Y}_v^{(f)}$. $\mathcal{F}_{\mathcal{B}_v}$ predicts the position-aware boundary maps $\hat{Y}_v^{(b)(p)} \in \mathbb{R}^{D \times T}$ and the channel-aware boundary maps $\hat{Y}_v^{(b)(c)} \in \mathbb{R}^{D \times T}$ as output. These results are aggregated by a convolutional layer which outputs position-channel boundary maps denoted as $\hat{Y}_v^{(b)(pc)} \in \mathbb{R}^{D \times T}$. Similarly, the audio boundary module $\mathcal{F}_{\mathcal{B}_a}$ input consists of the concatenation of latent features $\mathbf{z}_a$ and classification outputs $\hat{Y}_a^{(f)}$, i.e $\mathbf{z}_a \oplus \hat{Y}_a^{(f)}$. $\mathcal{F}_{\mathcal{B}_a}$ first predicts the audio position-aware boundary maps $\hat{Y}_a^{(b)(p)}$ and channel-aware boundary maps $\hat{Y}_a^{(b)(c)}$. Then $\hat{Y}_a^{(b)(p)}$ and $\hat{Y}_a^{(b)(c)}$ are aggregated to $\hat{Y}_a^{(b)(pc)}$ using a convolutional layer.

4.6. Multimodal fusion module

The fusion module illustrated in Fig. 5, uses boundary maps $\hat{Y}_v^{(b)(p)}$, $\hat{Y}_a^{(b)(p)}$, $\hat{Y}_v^{(b)(c)}$, $\hat{Y}_a^{(b)(c)}$, $\hat{Y}_v^{(b)(pc)}$, and $\hat{Y}_a^{(b)(pc)}$ and features $\mathbf{z}_v$ and $\mathbf{z}_a$ from the visual and audio modalities as input. Since the boundary module corresponding to each modality predicts three boundary maps, there are three fusion modules for position-aware $\mathcal{F}_{\Gamma_p}$, channel-aware $\mathcal{F}_{\Gamma_c}$ and aggregated position-channel $\mathcal{F}_{\Gamma_{pc}}$ boundary maps.

For the visual modality, the visual boundary maps and features from the visual and audio modalities are used to calculate the visual weights $W_v \in \mathbb{R}^{D \times T}$. Similarly, for the audio modality, the audio boundary maps and features from both modalities are utilized to calculate the audio weights $W_a \in \mathbb{R}^{D \times T}$. The element-wise weighted average of the fusion boundary maps predictions $\hat{Y}^{(b)(p)}$, $\hat{Y}^{(b)(c)}$ and $\hat{Y}^{(b)(pc)}$ is formed in the final step. Each boundary map $\alpha \in \{p, c, pc\}$ is calculated as follows,

$$\hat{Y}^{(b)(\alpha)} = \frac{W_v \hat{Y}_v^{(b)(\alpha)} + W_a \hat{Y}_a^{(b)(\alpha)}}{W_v + W_a},$$

where all operations are element-wise.

4.7. Loss functions

The training process of BA-TFD+ is guided by contrastive ($\mathcal{L}_c$), frame classification ($\mathcal{L}_f$), boundary matching ($\mathcal{L}_b$) and multimodal boundary matching ($\mathcal{L}_{bm}$) loss functions.

4.7.1. Contrastive loss

Contrastive loss has been proven to be helpful to eliminate the misalignment between different modalities (Chung and Zisserman, 2017; Chugh et al., 2020). Motivated by this, BA-TFD+ uses the latent visual and audio features $\mathbf{z}_v$ and $\mathbf{z}_a$ of real videos as positive pairs. On the other hand, latent features $\mathbf{z}_v$ and $\mathbf{z}_a$ with at least one modified modality are considered negative pairs (i.e. $Y^{(c)} = 0$). Thus, the contrastive loss minimizes the difference between the visual and audio modalities for positive pairs (i.e. $Y^{(c)} = 1$) and keeps that margin larger than δ for negative pairs. The contrastive loss is defined as follows,

$$\mathcal{L}_c = \frac{1}{C_f \sum \mathbb{T}} \sum_{i=1}^{n} Y_i^{(c)} d_i^2 + (1 - Y_i^{(c)}) \max(\delta - d_i, 0)^2$$

$$d_i = \|\mathbf{z}_{v_i} - \mathbf{z}_{a_i}\|_2,$$

where, n is the number of samples in the dataset, d_i is the ℓ_2 distance between visual and audio modality in the latent space, $Y_i^{(c)}$ is the label for contrastive learning and $\mathbb{T} = \{t_i\}_0^n$ where $\sum \mathbb{T}$ is the total number of frames in the dataset.

4.7.2. Frame classification loss

This is a standard frame level cross-entropy loss depicted as,

$$\mathcal{L}_f = -\frac{1}{2 \sum \mathbb{T}} \sum_{m \in \{a,v\}} \sum_{i=1}^{n} \sum_{j=1}^{t_i} H(\hat{Y}_{mij}^{(f)}, Y_{mij}^{(f)})$$

$$H(\hat{Y}^{(f)}, Y^{(f)}) = Y^{(f)} \log \hat{Y}^{(f)} + (1 - Y^{(f)}) \log (1 - \hat{Y}^{(f)})$$

$$Y_m^{(f)} = \eta_m Y^{(f)} + (1 - \eta_m) Y_\phi^{(f)},$$

where n is the number of samples in the dataset, t_i is the number of frames, m is the modality (i.e. audio a or visual v), η_m specifies whether modality m is manipulated or not, $Y_\phi^{(f)} \in 0^T$ is the label for real videos, and $\mathbb{T} = \{t_i\}_0^n$ where $\sum \mathbb{T}$ is the total number of frames in the dataset. This loss enforces the visual and audio encoder to learn whether a visual frame or audio sample is real or fake.

4.7.3. Boundary matching loss

Following the standard protocol (Lin et al., 2019; Su et al., 2021), we generated the ground truth boundary maps as labels for efficient training. The fusion boundary matching loss is calculated as,

$$\mathcal{L}_b = \frac{1}{3D \sum \mathbb{T}} \sum_{\alpha \in \{p,c,pc\}} \sum_{i=1}^{n} \sum_{j=1}^{D} \sum_{k=1}^{t_i} (\hat{Y}_{ijk}^{(b)(\alpha)} - Y_{ijk}^{(b)})^2,$$

where α is one of the boundary map types from the boundary module, n is the number of samples in the dataset, D is the maximum proposal duration, t_i is the number of frames, and $\mathbb{T} = \{t_i\}_0^n$ where $\sum \mathbb{T}$ is the total number of frames in the dataset.

[3] We only incorporate the multi-head self-attention layers of the ViT for the audio encoder.

Table 3
Temporal forgery localization results on the "fullset" of the LAV-DF dataset. The visual-only version of BA-TFD+ uses the output from the visual boundary matching layer, illustrating the performance when using only the visual modality.

Method	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
BMN (Lin et al., 2019)	10.56	01.66	00.00	48.49	44.39	37.13	31.55
BMN (E2E)	24.01	07.61	00.07	53.26	41.24	31.60	26.93
MDS (Chugh et al., 2020)	12.78	01.62	00.00	37.88	36.71	34.39	32.15
AGT (Nawhal and Mori, 2021)	17.85	09.42	00.11	43.15	34.23	24.59	16.71
BSN++ (Su et al., 2021)	56.41	32.57	00.21	74.93	71.11	64.98	59.29
AVFusion (Bagchi et al., 2022)	65.38	23.89	00.11	62.98	59.26	54.80	52.11
BA-TFD (Cai et al., 2022)	79.15	38.57	00.24	67.03	64.18	60.89	58.51
TadTR (Liu et al., 2022b)	80.22	61.04	**05.22**	72.50	72.50	70.56	69.18
ActionFormer (Zhang et al., 2022)	85.23	59.05	00.93	77.23	77.23	77.19	76.93
TriDet (Shi et al., 2023)	86.33	70.23	03.05	74.47	74.47	74.46	74.45
BA-TFD+ (ours)	**96.30**	**84.96**	04.44	**81.62**	**80.48**	**79.40**	**78.75**
BA-TFD+ (ours) (visual only)	64.78	54.85	02.53	64.00	59.33	55.94	54.38

Table 4
Temporal forgery localization results on the "subset" of the LAV-DF dataset. The visual-only version of BA-TFD+ uses the output from the visual boundary matching layer, illustrating the performance when using only the visual modality.

Method	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
BMN (Lin et al., 2019)	28.10	05.47	00.01	55.49	54.44	52.14	47.72
BMN (E2E)	32.32	11.38	00.14	59.69	48.17	39.01	34.17
MDS (Chugh et al., 2020)	23.43	03.48	00.00	58.53	56.68	53.16	49.67
AGT (Nawhal and Mori, 2021)	15.69	10.69	00.15	49.11	40.31	31.70	23.13
BSN++ (Su et al., 2021)	65.26	37.70	00.22	78.89	76.32	71.00	65.38
AVFusion (Bagchi et al., 2022)	62.01	22.77	00.11	61.98	58.08	53.31	50.52
BA-TFD (Cai et al., 2022)	85.20	47.06	00.29	67.34	64.52	61.19	59.32
TadTR (Liu et al., 2022b)	83.48	63.57	**05.44**	74.15	74.15	72.42	71.38
ActionFormer (Zhang et al., 2022)	79.48	48.01	01.08	70.38	70.38	70.36	70.08
TriDet (Shi et al., 2023)	80.71	60.93	02.91	67.64	67.64	67.64	67.63
BA-TFD+ (ours)	**96.82**	**86.47**	03.90	**81.74**	**80.59**	**79.60**	**79.15**
BA-TFD+ (ours) (visual only)	96.47	82.02	03.79	80.65	79.00	77.46	76.90

Table 5
Temporal forgery localization results on the ForgeryNet dataset. The visual-only version of BA-TFD+ uses the output from the visual boundary matching layer, illustrating the performance when using only the visual modality.

Method	Avg. AP	AP@0.5	AP@0.75	AP@0.95	AR@5	AR@2
Xception (Chollet, 2017)	62.83	68.29	62.84	58.30	73.95	25.83
X3D-M+BSN (Feichtenhofer, 2020; Lin et al., 2018)	70.29	80.46	77.24	55.09	86.88	81.33
X3D-M+BMN (Feichtenhofer, 2020; Lin et al., 2019)	83.47	90.65	88.12	74.95	91.99	88.44
SlowFast+BSN (Feichtenhofer et al., 2019; Lin et al., 2018)	73.42	82.25	80.11	60.66	88.78	83.63
SlowFast+BMN (Feichtenhofer et al., 2019; Lin et al., 2019)	86.85	92.76	**91.00**	80.02	93.49	**90.64**
BA-TFD+ (ours) (visual only)	**87.79**	**93.13**	89.14	**81.09**	**95.69**	90.63

4.7.4. Multimodal boundary matching loss

We utilized the label information for each modality to train the proposed multimodal framework and extended the concept of boundary matching loss ($\mathcal{L}_b$) to more modalities. The multimodal boundary matching loss is defined as follows,

$$\mathcal{L}_{bm} = \frac{1}{2D\sum\mathbb{T}} \sum_{m\in\{v,a\}} \sum_{\alpha\in\{p,c,pc\}} \sum_{i=1}^{n} \sum_{j=1}^{D} \sum_{k=1}^{t_i} (\hat{Y}^{(b)(\alpha)}_{mijk} - Y^{(b)}_{mijk})^2$$

$$Y^{(b)}_{m} = \eta_m Y^{(b)} + (1-\eta_m) Y^{(b)}_{\phi},$$

where, m is the modality (visual v or audio a), η_m specifies whether modality m is modified, α is one of the boundary map types from the boundary module, $Y^{(b)}_{\phi} \in 0^{D\times T}$ is the ground truth boundary maps for real videos, and $\mathbb{T} = \{t_i\}_0^n$ where $\sum\mathbb{T}$ is the total number of frames in the dataset.

4.7.5. Overall loss

The overall training objective of BA-TFD+ is defined as,

$$\mathcal{L} = \mathcal{L}_b + \lambda_{bm}\mathcal{L}_{bm} + \lambda_f\mathcal{L}_f + \lambda_c\mathcal{L}_c,$$

where, λ_{bm}, λ_f and λ_c are weights for different losses.

4.8. Inference

During inference, the model generates three types of fusion boundary maps - position-aware boundary map $\hat{Y}^{(b)(p)}$, channel-aware boundary map $\hat{Y}^{(b)(c)}$ and aggregated position-channel boundary map $\hat{Y}^{(b)(pc)}$. Following previous work (Su et al., 2021), we averaged the three boundary maps to produce the final boundary map $\hat{Y}^{(b)}$. This boundary map represents the confidence for all proposals in the video. Since this operation produces duplicated proposals, we post-process the proposals with Soft Non-Maximum Suppression (S-NMS) (Bodla et al., 2017) similar to BSN++ (Su et al., 2021).

5. Experiments

5.1. Dataset partitioning

We splitted the LAV-DF dataset into 78,703 train, 31,501 validation and 26,100 test videos. The test partition is denoted as *full set*. For a fair comparison with existing visual-only methods (Lin et al., 2019; Su et al., 2021), we additionally prepared a subset of the full set denoted as *subset* where the audio-only manipulated videos are removed.

Table 6
Deepfake detection results on the DFDC dataset.

Method	AUC
Meso4 (Afchar et al., 2018)	0.753
FWA (Li and Lyu, 2019)	0.727
Siamese (Mittal et al., 2020)	0.844
MDS (Chugh et al., 2020)	0.916
BA-TFD (Cai et al., 2022)	0.846
BA-TFD+ (ours)	**0.937**

5.2. Implementation details

The BA-TFD+ method is implemented in PyTorch (Paszke et al., 2019) and the model is trained using 2 NVIDIA A100 80 GB GPUs. We resized the input videos to 96 × 96 to reduce the computational cost of the MViTv2-based visual backbone. The temporal dimension T is fixed to 512 for LAV-DF and 300 for ForgeryNet (He et al., 2021) and DFDC (Dolhansky et al., 2020). The latent features $\mathbf{z}_v$ and $\mathbf{z}_a$ have the same shape $C_f \times T$ where the feature size $C_f = 256$ and $T \in \{512, 300\}$. For the boundary matching modules $\mathcal{F}_{\mathcal{B}_v}$ and $\mathcal{F}_{\mathcal{B}_a}$, we set the maximum segment duration D to 40 for LAV-DF, 200 for ForgeryNet and 300 for DFDC. We followed the training protocol proposed in MViTv2 (Li et al., 2022). Throughout our experiments, we empirically set $\lambda_c = 0.1$, $\lambda_f = 2$, $\lambda_b = 1$, $\lambda_{bm} = 1$ and $\delta = 0.99$.

5.3. Evaluation details

We benchmarked the LAV-DF dataset for deepfake detection and localization tasks. For deepfake detection we follow standard evaluation protocols (Rossler et al., 2019; Dolhansky et al., 2020), and use Area Under the Curve (AUC) as evaluation metric for this binary classification task. We are the first to benchmark deepfake localization task and adopt Average Precision (AP) and Average Recall (AR) as the evaluation metrics. For AP, we set the IoU thresholds to 0.5, 0.75 and 0.95, following ActivityNet (Caba Heilbron et al., 2015) evaluation protocol. For AR, since the number of fake segments is small, we set the number of proposals to 100, 50, 20 and 10 with the IoU thresholds [0.5:0.05:0.95]. When evaluating the proposed approach on ForgeryNet (He et al., 2021), we follow the protocol in that paper (i.e. AP@0.5, AP@0.75, AP@0.9, AR@5, and AR@2).

For evaluating BA-TFD+ on ForgeryNet, we used only the visual pipeline of the method to train the model (ForgeryNet is a visual-only deepfake dataset). Since only the visual modality is used in the model, only $\mathcal{L}_b$ and $\mathcal{L}_f$ are used for training. Similarly for evaluation on DFDC (Dolhansky et al., 2020), we consider the whole fake video as one fake segment and train our model in the temporal localization manner. Then, we train a small MLP to map the boundary map to the final binary labels.

We also evaluated the performance of several state-of-the-art methods on LAV-DF, including BMN (Lin et al., 2019), AGT (Nawhal and Mori, 2021), AVFusion (Bagchi et al., 2022), MDS (Chugh et al., 2020), BSN++ (Su et al., 2021), TadTR (Liu et al., 2022b), ActionFormer (Zhang et al., 2022), and TriDet (Shi et al., 2023). Based on the original implementations, BMN, BSN++, TadTR, ActionFormer, and TriDet require extracted features, thus, we trained these models based on 2-stream I3D features (Carreira and Zisserman, 2017). For the methods that require S-NMS (Bodla et al., 2017) during post-processing, we searched the optimal hyperparameters for S-NMS using the validation part of the concerned dataset. All reported results are based on the test partitions.

6. Results

6.1. Temporal forgery localization

6.1.1. LAV-DF dataset

We evaluated the performance of BA-TFD+ on the LAV-DF dataset for temporal forgery localization, and compare it with other approaches. For the full set, from Table 3, our method achieves the best performance for AP@0.5 and AR@100. Unlike temporal action localization datasets, the segments in our dataset have a single label for the fake segments which leads to high AP scores. The multimodal MDS method is not specifically designed for temporal forgery localization tasks and can only predict fixed-length segments, lacking the ability to precisely identify boundaries. Therefore, the scores for MDS are relatively low. For BMN and BSN++, the AP scores are low because they are designed for fake proposal generation instead of forgery localization. TadTR, ActionFormer, and TriDet achieve relatively better performance as they are one-stage temporal action localization approaches that generate more precise segments. Additionally, we observe that BMN trained with an end-to-end visual encoder performs better than using pre-trained I3D features. With the multimodal complimentary information, our approach outperforms the aforementioned approaches.

We further evaluated all methods on the subset of the LAV-DF dataset. From Table 4, it is observed that the performance of the visual-only methods including BMN, AGT, BSN++ and TadTR is improved. The visual-only score of our method improves from 64.78 (AP@0.5) to 96.47 (AP@0.5), and the margin between the unimodal and multimodal versions is decreased significantly from 31.52 (AP@0.5) to 0.35 (AP@0.5). Thus, our method demonstrates its superior performance for temporal forgery localization.

6.1.2. ForgeryNet dataset

We evaluated the performance of the visual-only BA-TFD+ trained on the ForgeryNet dataset, and compare it with other approaches (using the results reported by He et al. (2021)). As shown in Table 5, the performance of the visual-only BA-TFD+ exceeds the previous best model SlowFast (Feichtenhofer et al., 2019)+BMN (Lin et al., 2019), showing that proposed method has advantage for temporal forgery localization.

6.2. Deepfake detection

We also compare our method with previous deepfake detection methods on a subset of the DFDC dataset following the configuration of Chugh et al. (2020). As shown in Table 6, the performance of our method is better than previous methods such as Meso4 (Afchar et al., 2018), FWA (Li and Lyu, 2019), Siamese (Mittal et al., 2020), and MDS (Chugh et al., 2020). In summary, our method performs well on the classification task.

6.3. Ablation studies

6.3.1. Impact of loss functions

To examine the contributions of each loss of BA-TFD+, we train six models with different combinations of losses. To aggregate the frame-level predictions for the models without boundary module, we follow the algorithm proposed in previous work (Zhao et al., 2017). From Table 7, it is evident that all of the integrated losses have positive influence on the performance. By observing the difference between the scores, the boundary matching loss $\mathcal{L}_b$ and the frame classification loss $\mathcal{L}_c$ contribute significantly to the performance. With the frame-level labels supervising the model, the encoders are trained to have a better capacity to extract the features relevant to deepfake artifacts. Whereas

Table 7
Impact of loss functions. The contribution of different losses for temporal forgery localization on the full set of the LAV-DF dataset.

Loss function	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
$\mathcal{L}_f$	59.45	51.46	07.11	77.25	75.60	70.76	67.24
$\mathcal{L}_c, \mathcal{L}_f$	63.42	56.24	08.55	78.17	76.47	71.58	68.22
$\mathcal{L}_b$	71.31	34.30	00.12	66.92	63.67	57.99	54.72
$\mathcal{L}_{bm}, \mathcal{L}_b$	71.97	51.17	00.50	69.86	67.58	64.44	62.64
$\mathcal{L}_f, \mathcal{L}_{bm}, \mathcal{L}_b$	94.71	78.54	01.66	77.86	76.44	74.67	73.69
$\mathcal{L}_c, \mathcal{L}_f, \mathcal{L}_{bm}, \mathcal{L}_b$	**96.30**	**84.96**	**04.44**	**81.62**	**80.48**	**79.40**	**78.75**

Table 8
Impact of pre-trained features. Comparison of different pre-trained features for temporal forgery localization on the full set of the LAV-DF dataset.

Visual	Audio	Citation	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
I3D	E2E	Carreira and Zisserman (2017)	74.76	59.57	04.02	74.28	71.92	68.64	66.63
MARLIN	E2E	Cai et al. (2023)	92.27	75.11	04.10	77.93	76.38	74.53	73.47
3DMM	E2E	Blanz and Vetter (1999)	01.84	00.11	00.00	34.00	31.54	20.94	11.81
E2E	TRILLsson3	Shor and Venugopalan (2022)	95.16	82.67	05.65	81.21	79.80	78.22	77.49
E2E	Wav2Vec2	Baevski et al. (2020)	95.92	84.94	**05.66**	**82.48**	**81.38**	**79.93**	**79.24**
E2E	E2E	N/A	**96.30**	**84.96**	04.44	81.62	80.48	79.40	78.75

Table 9
Impact of encoder architectures. Comparison of different backbone architectures for temporal forgery localization on the full set of the LAV-DF dataset.

Visual	Audio	Boundary	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
3D CNN	CNN	BMN	76.90	38.50	00.25	66.90	64.08	60.77	58.42
3D CNN	CNN	BSN++	92.44	71.34	01.15	75.86	74.43	72.39	71.21
MViTv2-Tiny	CNN	BMN	89.32	59.47	01.45	72.52	70.14	67.55	65.92
MViTv2-Small	CNN	BMN	89.31	59.97	01.78	72.74	70.35	67.56	65.87
MViTv2-Base	CNN	BMN	89.90	59.67	01.51	72.22	69.99	67.29	65.64
3D CNN	ViT-Tiny	BMN	78.08	35.18	00.41	67.38	64.38	60.92	58.66
3D CNN	ViT-Small	BMN	79.61	37.63	00.42	67.10	64.23	60.77	58.51
3D CNN	ViT-Base	BMN	80.86	36.55	00.34	67.24	64.27	60.86	58.46
MViTv2-Small	ViT-Base	BSN++	93.59	75.22	02.56	77.73	76.08	74.07	72.93
MViTv2-Base	ViT-Base	BSN++	**96.30**	**84.96**	**04.44**	**81.62**	**80.48**	**79.40**	**78.75**

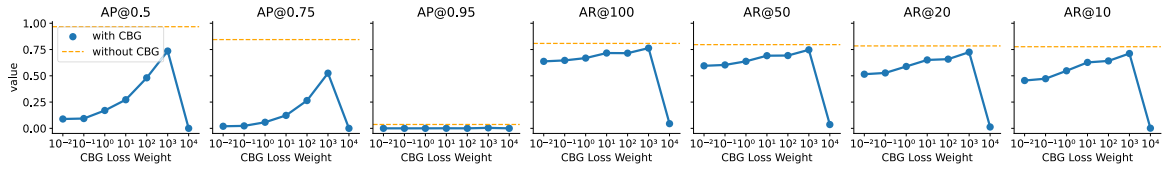


Fig. 6. Impact of CBG in the boundary matching module. The figure shows comparison of models containing CBG module and a model without CBG module.

the boundary module mechanism have localization ability to detect the fake segments more precisely.

6.3.2. Impact of pre-trained features

In the literature (Liu et al., 2022b; Zhang et al., 2022), pre-trained visual features, such as I3D (Carreira and Zisserman, 2017), are commonly used for temporal action localization. Since the I3D features are pre-trained on the Kinetics dataset (Kay et al., 2017), they encode the representation of the universal scene of the video. However, temporal forgery localization requires the model to have a specialized understanding of facial information. Therefore, the pre-trained features obtained from universal visual dataset are not likely to be suitable for our task. Our quantitative results support this, e.g. the comparison between the two BMN models in Table 3 where one uses I3D features and the other uses end-to-end training.

To examine the impact of pre-trained features on BA-TFD+, we trained models using different pre-trained features, including visual (I3D, MARLIN ViT-S Cai et al., 2023 and 3DMM Blanz and Vetter, 1999) and audio features (TRILLsson Shor and Venugopalan, 2022 and Wav2Vec2 Baevski et al., 2020). The results are shown in Table 8. From the results, we can observe the following patterns: (1) The model trained fully end-to-end reaches the best performance and (2) Compared with visual features, audio features have better task specific performance.

6.3.3. Impact of encoder architectures

To find the best modality-specific architecture for BA-TFD+, we trained several architecture combinations for the visual encoder, audio encoder, and boundary module. The results are presented in Table 9. Compared to the previous model BA-TFD (Cai et al., 2022) as baseline (3D-CNN + CNN + BMN Lin et al., 2019), we used the attention-based architectures including MViTv2 (Li et al., 2022) and ViT (Dosovitskiy et al., 2021) families for encoders and attention-based BSN++ modules (Su et al., 2021) for predicting boundaries.

We used the variations of MViTv2 from the original paper (i.e. MViTv2-Tiny, MViTv2-Small and MViTv2-Base) as the visual encoders. We can conclude that the MViTv2 architecture plays an important role while comparing with the baseline, but the benefit of different scales of the MViTv2 architecture is not significant. As for the audio encoder, we followed the architecture definitions for ViT (i.e. ViT-Tiny, ViT-Small and ViT-Base) for comparison. We can conclude that the audio encoder benefits from different scales of the ViT architecture. We also compared the BSN++-based boundary module with BMN-based architecture. The contribution from the BSN++ is the most significant compared with MViTv2 for the visual encoder and ViT for the audio encoder. Owing to the attention mechanism, the framework utilizes the global and local context to analyze the artifacts. The combination of MViTv2-Base, ViT-Base and BSN++ produces the best performance compared to all other combinations of modules.

6.3.4. Impact of CBG in the boundary matching module

We adopted the method from BSN++ (Su et al., 2021) to improve the performance for temporal forgery localization. This method includes two modules, complementary boundary generator (CBG) and proposal relation block (PRB). The CBG module predicts the confidence that a timestamp is starting or ending point of segments. The PRB module, based on BMN (Lin et al., 2019), predicts the boundary map which contains the confidences of dense segment proposals. For inference, the results from both modules are multiplied as the final output. In this ablation study, we aim to discuss the impact of the CBG module.

We trained several models containing CBG modules with different loss weights, from 10^{-2} to 10^{4}, and also a model without CBG module. As shown in Fig. 6, the best CBG loss weight is 10^{3}. However, compared with the non-CBG model, the best model with CBG can only compete on AR and has a huge gap on AP metrics. Based on this observation, we drop the CBG module in the boundary module and only use PRB.

7. Conclusion

In this paper, we introduce and investigate content-driven multimodal deepfake generation, detection, and localization. We introduce a new dataset where both the audio and visual modalities are modified at strategic locations. Additionally, we propose a new method for temporal forgery localization. Through extensive experiments, we demonstrate that our method outperforms existing state-of-the-art techniques.

The proposed dataset, LAV-DF, may raise ethical concerns due to its potential negative social impact. Given that the dataset contains facial videos of celebrities, there could be a risk of its misuse for unethical purposes. Moreover, the dataset generation pipeline itself can be used to generate fake videos. To confront the potential negative impact of our work, we have taken several measures. Most importantly, we have prepared an end-user license agreement as a preventive measure. Similarly, users need to agree on terms and conditions to use the proposed temporal forgery localization method BA-TFD+.

This work has some limitations: (1) The audio reenactment method employed for dataset creation does not consistently generate the desired reference style, (2) The resolution of the dataset is limited by the source videos, and (3) The high classification scores obtained indicate the need for further improvement in the visual reenactment method.

Major improvement in the future will be extending the generation pipeline to include word tokens insertion, substitution and deletion and converting statements into questions.

CRediT authorship contribution statement

Zhixi Cai: Methodology, Software, Validation, Data curation, Writing – original draft, Visualization. **Shreya Ghosh:** Methodology, Writing – original draft. **Abhinav Dhall:** Conceptualization, Resources, Supervision. **Tom Gedeon:** Writing – review & editing, Supervision. **Kalin Stefanov:** Writing – review & editing, Supervision. **Munawar Hayat:** Resources, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared the code and data in the github repository mentioned in the abstract.

References

Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., 2018. MesoNet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–7. http://dx.doi.org/10.1109/WIFS.2018.8630761, ISSN: 2157-4774.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, Vol. 33. Curran Associates, Inc., pp. 12449–12460.

Bagchi, A., Mahmood, J., Fernandes, D., Sarvadevabhatla, R., 2022. Hear me out: Fusional approaches for audio augmented temporal action localization:. In: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS - Science and Technology Publications, pp. 144–154. http://dx.doi.org/10.5220/0010832700003124.

Bayar, B., Stamm, M.C., 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. In: IH&MMSec '16, Association for Computing Machinery, New York, NY, USA, pp. 5–10. http://dx.doi.org/10.1145/2909827.2930786.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. KGIbfiiP1i4C, O'Reilly Media, Inc., Google-Books-ID.

Blanz, V., Vetter, T., 1999. A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99. ACM Press, Not Known, pp. 187–194. http://dx.doi.org/10.1145/311535.311556.

Bodla, N., Singh, B., Chellappa, R., Davis, L.S., 2017. Soft-NMS – improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5561–5569.

Brandon, J., 2019. There are now 15,000 deepfake videos on social media. yes, you should worry. Forbes.

Buch, S., Escorcia, V., Ghanem, B., Fei-Fei, L., Niebles, J.C., 2019. End-to-end, single-stream temporal action detection in untrimmed videos. In: Procedings of the British Machine Vision Conference 2017. Publisher: British Machine Vision Association, http://dx.doi.org/10.5244/c.31.93.

Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J., 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970.

Cai, Z., Ghosh, S., Stefanov, K., Dhall, A., Cai, J., Rezatofighi, H., Haffari, R., Hayat, M., 2023. MARLIN: Masked autoencoder for facial video representation LearnINg. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1493–1504.

Cai, Z., Stefanov, K., Dhall, A., Hayat, M., 2022. 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. http://dx.doi.org/10.1109/DICTA56598.2022.10034605.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308.

Casanova, E., Shulby, C., Gölge, E., Müller, N.M., Oliveira, F.S.D., Candido Jr., A., Soares, A.D.S., Aluisio, S.M., Ponti, M.A., 2021. SC-GlowTTS: An efficient zero-shot multi-speaker text-to-speech model. In: Interspeech 2021. ISCA, pp. 3645–3649. http://dx.doi.org/10.21437/Interspeech.2021-1774.

Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., Xu, C., 2020. Talking-head generation with rhythmic head motion. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 35–51. http://dx.doi.org/10.1007/978-3-030-58545-7_3.

Chen, B., Li, T., Ding, W., 2022. Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM. Inform. Sci. 601, 58–70. http://dx.doi.org/10.1016/j.ins.2022.04.014.

Chen, L., Maddox, R.K., Duan, Z., Xu, C., 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7832–7841.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Chugh, K., Gupta, P., Dhall, A., Subramanian, R., 2020. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, Association for Computing Machinery, New York, NY, USA, pp. 439–447. http://dx.doi.org/10.1145/3394171.3413700.

Chung, J.S., Nagrani, A., Zisserman, A., 2018. VoxCeleb2: Deep speaker recognition. In: Interspeech 2018. ISCA, pp. 1086–1090. http://dx.doi.org/10.21437/Interspeech.2018-1929.

Chung, J.S., Zisserman, A., 2017. Out of time: Automated lip sync in the wild. In: Chen, C.-S., Lu, J., Ma, K.-K. (Eds.), Computer Vision – ACCV 2016 Workshops. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 251–263. http://dx.doi.org/10.1007/978-3-319-54427-4_19.

Coccomini, D.A., Messina, N., Gennaro, C., Falchi, F., 2022. Combining EfficientNet and vision transformers for video deepfake detection. In: Sclaroff, S., Distante, C., Leo, M., Farinella, G.M., Tombari, F. (Eds.), Image Analysis and Processing – ICIAP 2022. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 219–229. http://dx.doi.org/10.1007/978-3-031-06433-3_19.

Cozzolino, D., Poggi, G., Verdoliva, L., 2017. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. In: IH&MMSec '17, Association for Computing Machinery, New York, NY, USA, pp. 159–164. http://dx.doi.org/10.1145/3082031.3083247.

Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M., 2022. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. Int. J. Comput. Vis. 130 (1), 33–55. http://dx.doi.org/10.1007/s11263-021-01531-2.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C., 2020. The DeepFake detection challenge (DFDC) dataset. arXiv:2006.07397 [cs].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Feichtenhofer, C., 2020. X3D: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–213.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. SlowFast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6202–6211.

Fellbaum, C., 1998. WordNet: an Electronic Lexical Database. MIT Press, Google-Books-ID: Rehu8OOzMIMC.

Gao, J., Chen, K., Nevatia, R., 2018. CTAP: Complementary temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 68–83.

Gao, J., Yang, Z., Chen, K., Sun, C., Nevatia, R., 2017. TURN TAP: Temporal unit regression network for temporal action proposals. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3628–3636.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144. http://dx.doi.org/10.1145/3422622.

Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., Ma, L., 2021. Spatiotemporal inconsistency learning for DeepFake video detection. In: Proceedings of the 29th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, pp. 3473–3481.

Guarnera, L., Giudice, O., Battiato, S., 2020. DeepFake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 666–667.

Guo, Y., Chen, K., Liang, S., Liu, Y.-J., Bao, H., Zhang, J., 2021. AD-NeRF: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794.

He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z., 2021. ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4360–4369.

Heo, Y.-J., Yeo, W.-H., Kim, B.-G., 2023. DeepFake detection algorithm based on improved vision transformer. Appl. Intell. 53 (7), 7512–7527. http://dx.doi.org/10.1007/s10489-022-03867-9.

Idrees, H., Zamir, A.R., Jiang, Y.-G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M., 2017. The THUMOS challenge on action recognition for videos "in the wild". Comput. Vis. Image Underst. 155, 1–23. http://dx.doi.org/10.1016/j.cviu.2016.10.018, arXiv:1604.06182.

Ilyas, H., Javed, A., Malik, K.M., 2023. AvFakeNet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection. Appl. Soft Comput. 136, 110124. http://dx.doi.org/10.1016/j.asoc.2023.110124.

Jamaludin, A., Chung, J.S., Zisserman, A., 2019. You said that?: Synthesising talking faces from audio. Int. J. Comput. Vis. 127 (11), 1767–1779. http://dx.doi.org/10.1007/s11263-019-01150-y.

Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I.L., Wu, Y., 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS '18, Curran Associates Inc., Red Hook, NY, USA, pp. 4485–4495.

Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C., 2020. 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.

K R, P., Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V., Jawahar, C.V., 2019. Towards automatic face-to-face translation. In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19, Association for Computing Machinery, New York, NY, USA, pp. 1428–1436. http://dx.doi.org/10.1145/3343031.3351066.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A., 2017. The kinetics human action video dataset. http://dx.doi.org/10.48550/arXiv.1705.06950, arXiv:1705.06950 [cs].

Khalid, H., Kim, M., Tariq, S., Woo, S.S., 2021a. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In: Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection. pp. 7–15. http://dx.doi.org/10.1145/3476099.3484315, arXiv:2109.02993.

Khalid, H., Tariq, S., Woo, S.S., 2021b. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. arXiv:2108.05080 [cs].

King, D.E., 2009. Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res. 10, 1755–1758.

Korshunov, P., Marcel, S., 2018. DeepFakes: a new threat to face recognition? Assessment and detection. arXiv:1812.08685 [cs].

Korshunova, I., Shi, W., Dambre, J., Theis, L., 2017. Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3677–3685.

Kwon, P., You, J., Nam, G., Park, S., Chae, G., 2021. KoDF: A large-scale Korean DeepFake detection dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10744–10753.

Lewis, J.K., Toubal, I.E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z., Prasad, C., Palaniappan, K., 2020. Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In: 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). pp. 1–9. http://dx.doi.org/10.1109/AIPR50011.2020.9425167, ISSN: 2332-5615.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B., 2020a. Face X-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001–5010.

Li, Y., Lyu, S., 2019. Exposing DeepFake videos by detecting face warping artifacts. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). p. 7.

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C., 2022. MViTv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020b. 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.

de Lima, O., Franklin, S., Basu, S., Karwoski, B., George, A., 2020. Deepfake detection using spatiotemporal convolutional networks. arXiv:2006.14749 [cs, eess].

Lin, T., Liu, X., Li, X., Ding, E., Wen, S., 2019. BMN: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3889–3898.

Lin, T., Zhao, X., Shou, Z., 2017. Single shot temporal action detection. In: Proceedings of the 25th ACM International Conference on Multimedia. MM '17, Association for Computing Machinery, New York, NY, USA, pp. 988–996. http://dx.doi.org/10.1145/3123266.3123343.

Lin, T., Zhao, X., Su, H., Wang, C., Yang, M., 2018. BSN: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19.

Liu, X., Bai, S., Bai, X., 2022a. An empirical study of end-to-end temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20010–20019.

Liu, X., Hu, Y., Bai, S., Ding, F., Bai, X., Torr, P.H.S., 2021. Multi-shot temporal event localization: A benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12596–12606.

Liu, X., Wang, Q., Hu, Y., Tang, X., Zhang, S., Bai, S., Bai, X., 2022b. End-to-end temporal action detection with transformer. IEEE Trans. Image Process. 31, 5427–5441. http://dx.doi.org/10.1109/TIP.2022.3195321, Conference Name: IEEE Transactions on Image Processing.

Liu, Y., Wang, L., Wang, Y., Ma, X., Qiao, Y., 2022c. FineAction: A fine-grained video dataset for temporal action localization. IEEE Trans. Image Process. 31, 6937–6950. http://dx.doi.org/10.1109/TIP.2022.3217368, Conference Name: IEEE Transactions on Image Processing.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D., 2020. Emotions don't Lie: An audio-visual deepfake detection method using affective cues. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, Association for Computing Machinery, New York, NY, USA, pp. 2823–2832. http://dx.doi.org/10.1145/3394171.3413570.

Montserrat, D.M., Hao, H., Yarlagadda, S.K., Baireddy, S., Shao, R., Horvath, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., Delp, E.J., 2020. Deepfakes detection with automatic face weighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 668–669.

Narayan, K., Agarwal, H., Thakral, K., Mittal, S., Vatsa, M., Singh, R., 2023. DF-Platter: Multi-face heterogeneous deepfake dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9739–9748.

Nawhal, M., Mori, G., 2021. Activity graph transformer for temporal action localization. arXiv:2101.08540 [cs].

Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F., McAuley, J., 2021. Expressive neural voice cloning. In: Proceedings of the 13th Asian Conference on Machine Learning. PMLR, pp. 252–267, ISSN: 2640-3498.

Nick, D., Andrew, J., 2019. Contributing data to deepfake detection research.

Nirkin, Y., Keller, Y., Hassner, T., 2019. FSGAN: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7184–7193.

Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. WaveNet: A generative model for raw audio. arXiv:1609.03499 [cs].

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, Vol. 32. Curran Associates, Inc..

Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C., 2020. A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, Association for Computing Machinery, New York, NY, USA, pp. 484–492. http://dx.doi.org/10.1145/3394171.3413532.

Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J., 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 86–103. http://dx.doi.org/10.1007/978-3-030-58610-2_6.

Raza, M.A., Malik, K.M., 2023. Multimodaltrace: Deepfake detection using audiovisual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 993–1000.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M., 2019. FaceForensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1985. Learning Internal Representations by Error Propagation. Technical Report, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE, Section: Technical Reports.

Sample, I., 2020. What are deepfakes – and how can you spot them? Guardian.

Sanderson, C. (Ed.), 2002. The VidTIMIT Database. IDIAP.

Schwartz, O., 2018. You thought fake news was bad? Deep fakes are where truth goes to die. Guardian.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R.A., Agiomvrgiannakis, Y., Wu, Y., 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4779–4783. http://dx.doi.org/10.1109/ICASSP.2018.8461368, ISSN: 2379-190X.

Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D., 2023. TriDet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18857–18866.

Shor, J., Venugopalan, S., 2022. TRILLsson: Distilled universal paralinguistic speech representations. In: Interspeech 2022. pp. 356–360. http://dx.doi.org/10.21437/Interspeech.2022-118, arXiv:2203.00236 [cs, eess].

Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.-F., 2017. CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5734–5743.

Shou, Z., Wang, D., Chang, S.-F., 2016. Temporal action localization in untrimmed videos via multi-stage CNNs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1049–1058.

Su, H., Gan, W., Wu, W., Qiao, Y., Yan, J., 2021. BSN++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. pp. 2602–2610. http://dx.doi.org/10.1609/aaai.v35i3.16363, Number: 3.

Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M., 2020. Neural voice puppetry: Audio-driven facial reenactment. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), ECCV 2020. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 716–731. http://dx.doi.org/10.1007/978-3-030-58517-4_42.

Thomas, D., 2020. Deepfakes: A threat to democracy or just a bit of fun? BBC News.

Tulyakov, S., Liu, M.-Y., Yang, X., Kautz, J., 2018. MoCoGAN: Decomposing motion and content for video generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1526–1535.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., Saurous, R.A., 2017. Tacotron: Towards end-to-end speech synthesis. In: Interspeech 2017. ISCA, pp. 4006–4010. http://dx.doi.org/10.21437/Interspeech.2017-1452.

Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., Li, S.-N., 2022. M2TR: Multi-modal multi-scale transformers for deepfake detection. In: Proceedings of the 2022 International Conference on Multimedia Retrieval. ICMR '22, Association for Computing Machinery, New York, NY, USA, pp. 615–623. http://dx.doi.org/10.1145/3512527.3531415.

Wodajo, D., Atnafu, S., 2021. Deepfake video detection using convolutional vision transformer. arXiv:2102.11126 [cs].

Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B., 2020. G-TAD: Sub-graph localization for temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10156–10165.

Yang, M., Chen, G., Zheng, Y.-D., Lu, T., Wang, L., 2023a. BasicTAD: An astounding RGB-only baseline for temporal action detection. Comput. Vis. Image Underst. 232, 103692. http://dx.doi.org/10.1016/j.cviu.2023.103692.

Yang, X., Li, Y., Lyu, S., 2019. Exposing deep fakes using inconsistent head poses. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8261–8265. http://dx.doi.org/10.1109/ICASSP.2019.8683164, ISSN: 2379-190X.

Yang, K., Qiao, P., Li, D., Lv, S., Dou, Y., 2018. Exploring temporal preservation networks for precise temporal action localization. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32. Number: 1.

Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., Cao, X., Ren, K., 2023b. AVoiD-DF: Audio-visual joint learning for detecting deepfake. IEEE Trans. Inf. Forensics Secur. 18, 2015–2029. http://dx.doi.org/10.1109/TIFS.2023.3262148, Conference Name: IEEE Transactions on Information Forensics and Security.

Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C., 2019. Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7094–7103.

Zhang, C.-L., Wu, J., Li, Y., 2022. ActionFormer: Localizing moments of actions with transformers. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022. In: Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, pp. 492–510. http://dx.doi.org/10.1007/978-3-031-19772-7_29.

Zhao, H., Torralba, A., Torresani, L., Yan, Z., 2019. HACS: Human action clips and segments dataset for recognition and temporal localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8668–8678.

Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D., 2017. Temporal action detection with structured segment networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2914–2923.

Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D., 2020. MakeltTalk: speaker-aware talking-head animation. ACM Trans. Graph. 39 (6), 221:1–221:15. http://dx.doi.org/10.1145/3414685.3417774.

Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z., 2021a. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4176–4186.

Zhou, T., Wang, W., Liang, Z., Shen, J., 2021b. Face forensics in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5778–5788.

Zhu, Y., Gao, J., Zhou, X., 2023. AVForensics: Audio-driven deepfake video detection with masking strategy in self-supervision. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. ICMR '23, Association for Computing Machinery, New York, NY, USA, pp. 162–171. http://dx.doi.org/10.1145/3591106.3592218.

Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.-G., 2020. WildDeepfake: A challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, Association for Computing Machinery, New York, NY, USA, pp. 2382–2390. http://dx.doi.org/10.1145/3394171.3413769.