"Glitch in the Matrix!": A Large Scale Benchmark for Content Driven Audio-Visual Forgery Detection and Localization

Zhixi Cai[1], Shreya Ghosh[2], Tom Gedeon[2], Abhinav Dhall[1,3], Kalin Stefanov[1], Munawar Hayat[1]

[1]Monash University, [2] Curtin University, [3] Indian Institute of Technology Ropar,

{zhixi.cai,kalin.stefanov,munawar.hayat}@monash.edu,

{shreya.ghosh,tom.gedeon}@curtin.edu.au,abhinav@iitrpr.ac.in

Abstract

Most deepfake detection methods focus on detecting spatial and/or spatio-temporal changes in facial attributes. This is because available benchmark datasets contain mostly visual-only modifications. However, a sophisticated deepfake may include small segments of audio or audio-visual manipulations that can completely change the meaning of the content. To addresses this gap, we propose and benchmark a new dataset, Localized Audio Visual Deep-Fake (LAV-DF), consisting of strategic content-driven audio, visual and audio-visual manipulations. The proposed baseline method, Boundary Aware Temporal Forgery De-tection (BA-TFD), is a 3D Convolutional Neural Network-based architecture which efficiently captures multimodal manipulations. We further improve (i.e. BA-TFD+) the baseline method by replacing the backbone with a Multi-scale 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

[1]Recent advances in computer vision and image processing methods (*e.g.* Autoencoders [69] and Generative Adversarial Networks [29]) have enabled the creation of highly realistic fake videos, termed as *deepfakes*[2]. Deepfake generation methods mainly include voice cloning [38, 83], face reenactment [67, 81], and face swapping [46, 64]. Highly realistic deepfakes are a potential tool for spreading harm-ful misinformation given our increasing online presence. Therefore, this success in generating high-quality deepfakes has raised serious concerns about their role in shaping people's beliefs, with some scholars suggesting that deepfakes are a "threat to democracy" [7,70,72,80]. As an example of the potentially harmful effect of deepfakes, consider the recent work [79] that uses a video of the former United States President Barack Obama to showcase a novel face reenactment method. In this work, the lip movements of Barack Obama are synchronized with another person's speech, resulting in high-quality and realistic video in which the former president appears to say something he never did.

Given the recent surge in fake video content on the Internet, it has become increasingly important to identify deepfakes with more accurate and reliable methods. This has led to the release of several benchmark datasets [22,45,68] and methods [58] for fake content detection. These fake video detection methods aim to correctly classify any given input video as either *real* or *fake*. This means that the main assumption behind those datasets and methods is that fake content is present in the entirety of the visual/audio signal; that is, there is some form of manipulation of the content. Current state-of-the-art deepfake detection methods [20, 34, 84] achieve impressive results on this binary classification problem.

However, fake content might constitute only a small part of an otherwise long real video, as was initially suggested in [17]. Such short modified segments have the power to alter the meaning and sentiment of the original content completely. For example, consider the manipulation illustrated in Figure 1. The real video might represent a person saying "Vaccinations are safe", while the fake includes only a short modified segment; for example, "safe" is replaced with "dangerous". Hence, the meaning and sentiment of the fake video differ significantly from the real one. If done precisely, this type of coordinated manipulation can sway public opinion (*e.g.* when employed for media of a famous person as the example with Barack Obama) in a particular direction, for example, based on target sentiment polar-

---

[1]The paper is under consideration/review at Computer Vision and Image Understanding Journal.

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

ity. Given the discussed central assumption behind current datasets and methods, the state-of-the-art deepfake detectors might not perform well on this type of manipulations.

This paper tackles the important task of detecting content altering fake segments in videos. The literature review on benchmark datasets for deepfake detection indicates that there is no dataset suitable for this task, that is, a dataset that consists of content-driven manipulations. Therefore, this paper describes the process of creating such a large-scale dataset that will enable further research in this direction. In addition, the paper proposes a novel multimodal method for precisely detecting the boundaries of fake segments based on visual and audio information. The **main contributions** of our work are as follows,

- We introduce a new large-scale public dataset, *Localized Audio Visual DeepFake*, for temporal forgery localization.

- We propose a new method, *Boundary Aware Temporal Forgery Detection Plus*, for temporal forgery localization.

- We present a thorough experimental validation of the method's components and provide a comprehensive comparison with the state-of-the-art.

2. Related Work

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

2.1. Deepfake Datasets

The body of research in deepfake detection is driven by seminal datasets curated with different manipulation methods. A summary of the relevant datasets is presented in Table 1. [45] curated one of the first deepfake datasets, DF-TIMIT, where face swapping was performed on VidTimit [71]. Later, other important datasets such as UADFV [86], FaceForensics++ [68], and Google DFD [63] were introduced. Due to the complexity of face manipulation and limited availability of open-source face manipulation techniques, these datasets are fairly small in size [51]. Facebook released a large-scale dataset DFDC [22] in 2020 for the task of deepfake classification. Multiple face manipulation methods generated 128,154 videos, including real videos of 3,000 actors. DFDC has become a mainstream benchmark dataset for the task of deepfake detection. With the progress in both audio and visual deepfake manipulation, post DFDC, several new datasets including Celeb-DF [51], DeeperForensics [39], and WildDeepfake [94] were introduced. All these datasets are designed for the binary task of deepfake classification and focus primarily on

Figure 1. **Content-driven audio-visual manipulation.** Here, we display an example of content driven forgery. The left one is a real video with the subject saying "Vaccinations are safe". Right one is an audio-visual deepfake created from the real video based on the change in perceived sentiment where "safe" is changed to "dangerous". Green-edge and red-edge images are real and fake frames, respectively. *Please note that through subtle audio-visual manipulation, the whole meaning of the video content has changed.*

visual manipulation detection [17]. In 2021, OpenForensics [47] dataset was introduced for spatial detection, segmentation and classification. Recently, FakeAVCeleb [43] was released, focusing on both face swapping and face reenactment methods with manipulated audio and visual modalities. ForgeryNet [33] is the latest contribution to the growing list of deepfake detection datasets. This large-scale dataset is also centered around visual-only identity manipulations and is suitable for video/image classification and spatial/temporal forgery localization tasks.

Previous datasets include face manipulations that occur in most of the visual frames [17]. Only the latest one, ForgeryNet, provides examples of the problem of temporal forgery localization since it includes random face swapping applied to parts of some videos. However, the manipulations present in ForgeryNet are only identity modifications that do not necessarily alter the meaning of the content. Our content-driven manipulation dataset addresses this gap.

2.2. Deepfake Detection

Deepfake detection methods draw inspiration from observations of artifacts such as different eye colors, unnatural blinks and lip-sync issues in deepfake videos. These binary classification approaches are based on both traditional machine learning methods (*e.g.* EM [31] and SVM [88]) and deep learning methods (*e.g.* 3D-CNN [21], GRU [60] and ViT [20, 34, 84]). Previous methods [30, 48] also aim to detect temporal inconsistencies in deepfake content and recently, several audio-visual deepfake detection methods such as MDS [17] and M2TR [82] were proposed.

The methods above are classification centric and do not focus on temporal localization. The only exception is MDS, shown to work for localization tasks, however, the method was designed primarily for classification. We propose a dataset and method that are specifically designed for temporal localization of audio/visual and audio-visual manipulations.

Table 1. **Comparison of the proposed dataset with other publicly available deepfake datasets.** *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 [45]	2018	Cla	V	FS	43	320	640	960
UADFV [88]	2019	Cla	V	FS	49	49	49	98
FaceForensics++ [68]	2019	Cla	V	FS/RE	-	1,000	4,000	5,000
Google DFD [63]	2019	Cla	V	FS	-	363	3,068	3,431
DFDC [22]	2020	Cla	AV	FS	960	23,654	104,500	128,154
DeeperForensics [39]	2020	Cla	V	FS	100	50,000	10,000	60,000
Celeb-DF [51]	2020	Cla	V	FS	59	590	5,639	6,229
WildDeepfake [94]	2021	Cla	-	-	-	3,805	3,509	7,314
FakeAVCeleb [43]	2021	Cla	AV	RE	600+	570	25,000+	25,500+
ForgeryNet [33]	2021	SL/TFL/Cla	V	Random FS/RE	5,400+	99,630	121,617	221,247
LAV-DF (ours)	2022	TFL/Cla	AV	Content-driven RE	153	36,431	99,873	136,304

2.3. Temporal Action Localization

Given that the task of temporal forgery localization is similar to the task of temporal action localization, previous work in this area is important. Benchmark datasets in this domain include THUMOS [35] and ActivityNet [9] and the proposed methods can be grouped into two categories: 2-step approaches which first generate segment proposals and then perform multi-class classification to evaluate the proposals [56, 85, 89] and 1-step approaches which directly generate the final segment predictions [8, 53, 55, 61, 90]. For temporal forgery localization, there are no classification requirements for the foreground segments; the background is always real, and the foreground segments are always fake. Therefore, boundary prediction and 1-step approaches are more relevant for this task. [3] divided the approaches to segment proposal estimation in temporal action localization into two main categories: methods based on anchors and methods based on predicting the boundary probabilities. As for the anchor-based, these methods mainly use sliding windows in the video, such as S-CNN [77], CDC [76], TURN-TAP [28] and CTAP [27]. For the methods predicting the boundary probabilities, [54] introduced BSN. The method can utilize global information to overcome the problem that anchor-based methods cannot generate precise and flexible segment proposals. Based on BSN, BMN [52] and BSN++ [78] were introduced for improved performance. It is worth noting that all those methods are unimodal, which is not optimal for the task of temporal forgery localization. The importance of multimodality was recently demonstrated by AVFusion [3].

2.4. Proposed Approach

For the task of temporal forgery localization, both the audio and visual information are important, in addition to the required precise boundary proposals. In this paper, we introduce a multimodal method based on boundary probabilities and compare its performance with BMN [52],

AGT [61], MDS [17], AVFusion [3], BSN++ [78], TadTR [57], ActionFormer [90] and TriDet [74].

3. Localized Audio Visual DeepFake Dataset

We created a large-scale audio-visual deepfake dataset containing 136,304 video clips (36,431 real and 99,873 fake). The data generation pipeline is illustrated in Figure 2. Our deepfake generation method is based on changing relevant words in the video transcripts. Technically, our manipulation strategy includes replacing strategic words with their antonyms, which leads to a significant change in the perceived sentiment of a given statement.

3.1. Audio-Visual Data Sourcing

The real videos are sourced from the VoxCeleb2 dataset [18], a facial video dataset with over 1 million utterance videos of over 6,000 speakers. The faces in the videos are tracked and cropped with the Dlib facial detector [44] at 224×224 resolution. VoxCeleb2 includes variations in terms of duration, spoken language, and voice loudness. For simplicity, we choose English-speaking videos using the confidence score from the Google Speech-to-Text service[3]. Similarly, we use the same service to generate the video transcripts.

3.2. 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

After collecting and wrangling the real data from VoxCeleb2, the next step is to analyse a video's transcript de-

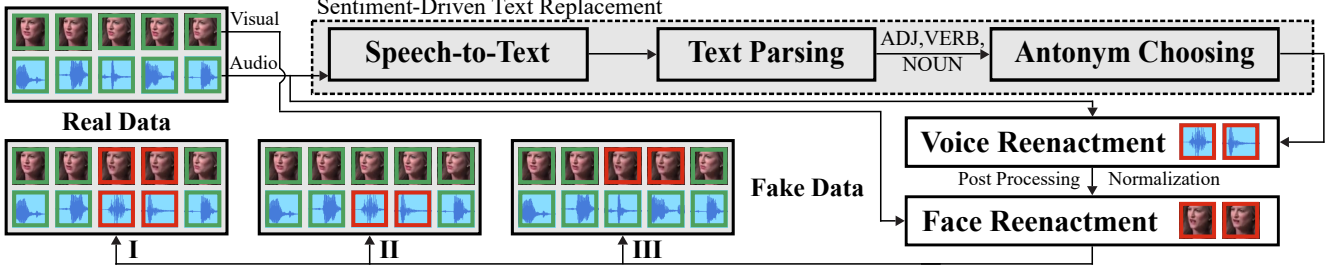---

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

Figure 2. **Data generation pipeline of LAV-DF**. The green-edge audio and visual frames are the real data, and the red-edge audio and visual frames are the generated fake data. The real audio-based transcript is used to decide the location and content to be replaced based on the largest change in sentiment. The chosen antonyms are used as input for generating fake audio with voice cloning. The post-processing and normalization are applied to the audio to maintain the consistency of the loudness between the generated audio and real audio in the neighborhood. The generated audio is used as input for facial reenactment. *Three categories of data are generated: Fake Audio and Fake Visual, Fake Audio and Real Visual and Real Audio and Fake Visual.*

noted by $D = \{d_0, d_1, \cdots, d_m, \cdots, d_n\}$, where d_i denotes word tokens and n is the number of tokens in the transcript. The aim is to find the tokens to be replaced in D such that the sentiment score of the transcript changes the most. This is essentially to create transcript $D' = \{d_0, d_1, \cdots, d'_m, \cdots, d_n\}$, composed of most of the tokens of D with the exception of a few tokens being replaced. These replaced tokens d' are selected from a set $\hat{d}$ of antonyms of d from WordNet [26]. We used the sentiment analyzer in NLTK [4] which predicts the sentiment value of a video transcript. For each token d in a transcript D, we find the replacement as follows,

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

We find all the replacements in a transcript D as,

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

where $\Delta S(\tau_i)$ is the sentiment difference with the replacement τ_i and M is the maximum number of replacements in the transcript.

There is up to 1 replacement for videos shorter than 10 seconds; otherwise, there are up to 2 replacements. Figure 3 (a) illustrates the change in sentiment distribution after the manipulations and Figure 3 (b) presents the histogram of $|\Delta S|$, suggesting that the sentiment of most transcripts was successfully changed.

3.2.2 Audio Generation

After the transcript manipulation, the next step is to generate speaker specific audio. Motivated by the prior work on adaptive text-to-speech (TTS) methods [13, 38, 62], we adopted SV2TTS [38] for person specific audio generation.

SV2TTS comprises three modules: 1) An encoder for extracting style embedding of the reference speaker, 2) A Tacotron 2-based [73] spectrogram generation using the replacement tokens and the speaker style embedding, and 3) A WaveNet-based [65] vocoder for generating realistic audio using the spectrogram. The pre-trained SV2TTS is used for generating the fake audio segments which are later loudness normalized using the corresponding real audio neighbors. The rationale behind the loudness normalization is to generate more realistic counterpart of the audio signal suitable for replacement.

3.2.3 Video Generation

The generated fake audio is further used as input for generating the corresponding fake visual frames. Wav2Lip [67] facial reenactment is used for this task as it has been shown to have state-of-the-art output generation quality along with better generalization capability and robustness to unseen scenarios [37,40]. We encountered several issues with using other popular generation methods such as AD-NeRF [32] and ATVGnet [15] in our generation context. For example, AD-NeRF is not designed for zero-shot generation of unseen identities, and ATVGnet reenacts the face based on a static reference image, which causes pose inconsistencies on the boundary between fake and real segments. Wav2Lip takes a reference video and target audio as input, and generates an output video in which the person in the reference video speaks the target audio content with synced lips. The pre-trained Wav2Lip is used and the generated fake visual segments are up-scaled to a resolution of 224×224. Most importantly, the generated fake audio and visual segments are synchronized and therefore can be used to replace the original audio and visual segments.

Similar to [42], the LAV-DB includes three categories of generated data.

- **Fake Audio** and **Fake Visual.** Both real audio and

visual segments corresponding to the replacement token(s) are manipulated.

- **Fake Audio** and **Real Visual.** Only the real audio corresponding to the replacement token(s) is manipulated. To keep the fake audio and real visual segments synchronized, the corresponding real visual segment is length-normalized.

- **Real Audio** and **Fake Visual.** Only the real visual segment corresponding to the replacement token(s) is manipulated and the length of the fake visual is normalized to match the length of the real audio.

3.3. Dataset Statistics

The dataset contains 136,304 videos, of which 36,431 are completely real, and 99,873 have fake segments, with 153 unique identities. We split the dataset into 3 identity-independent subsets for training (78,703 videos of 91 identities), validation (31,501 videos of 31 identities), and testing (26,100 videos of 31 identities). The summary of the dataset is shown in Figure 3. The total number of fake segments is 114,253, with duration in the range [0-1.6] seconds and an average length of 0.65 seconds, where 89.26% of the segments are shorter than 1 second. The maximum video length is 20 seconds, and 69.61% of the videos are shorter than 10 seconds. As for the modality modification types, the amount of the 4 types (*i.e.* visual-modified, audio-modified, both-modified, real) is approximately equal. In most videos (62.72%), there is 1 fake segment, and in some videos (10.55%), there are 2.

3.4. Data Generation Quality

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

Table 2. **Data generation comparison in terms of visual quality.** We maintained the experimental protocol and adopted the scores on VoxCeleb2 for the related deepfake generation methods from [92].

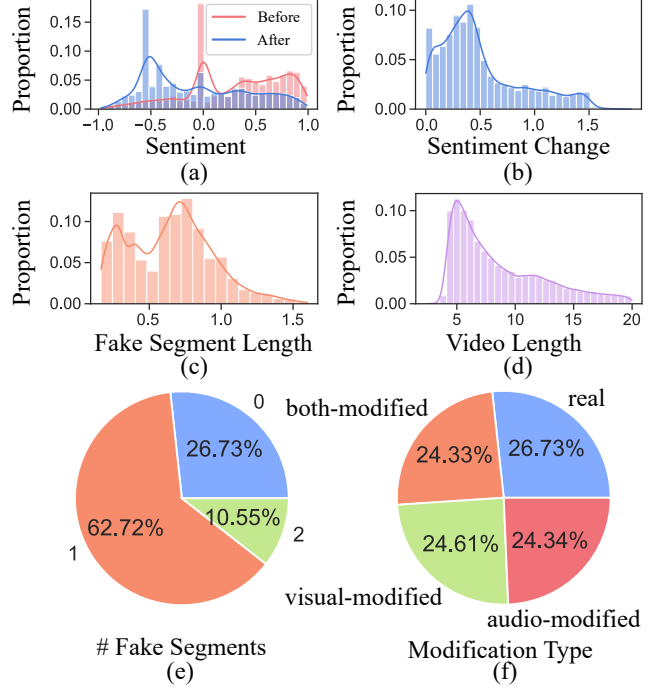| Method	PSNR	SSIM
ATVGnet [15]	29.41	0.826
Wav2Lip [67]	29.54	0.846
MakeitTalk [93]	29.51	0.817
Rhythmic Head [14]	29.55	0.779
PC-AVS [92]	29.68	0.886
LAV-DF (Ours)	**33.06**	**0.898**

Figure 3. **LAV-DB statistics.** (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.

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 multimodal segments. An overview of the proposed method is depicted in Figure 4 and Algorithm 1.

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.

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 RGB information $V \supset \{V_i\}_{i=1}^n$ using an MViTv2 [50]. MViTv2 achieves seminal performance gain for different video anal-
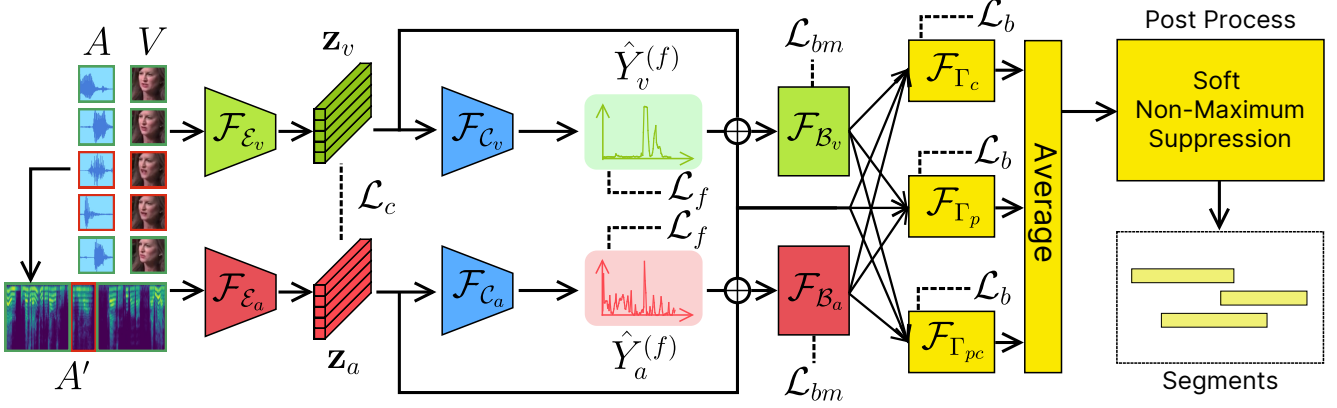
Figure 4. **Architectural overview of BA-TFD+ framework (best viewed in color).** BA-TFD+ mainly comprises of 1) Visual encoder ($\mathcal{F}_{\mathcal{E}_v}$) that takes resized raw video frames as input, 2) Audio encoder ($\mathcal{F}_{\mathcal{E}_a}$) that takes spectrogram extracted from raw audio 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.*
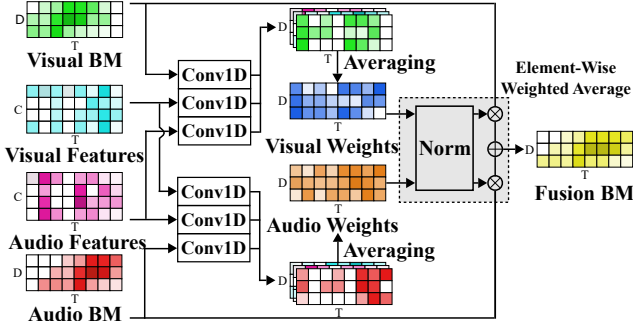

Figure 5. **Structure of the fusion module.** The gray block normalizes the visual and audio weights predicted from the 1D convolutional layers and applies element-wise weighted average. $\oplus$ *denotes element-wise addition,* $\otimes$ *denotes element-wise multiplication and BM denotes boundary map.*

ysis tasks including video action recognition and detection. In addition, MViTv2 leverages hierarchical multi-scale features compared to the basic ViT [23]. Our backbone MViTv2-Base model comprises of 4 blocks and 24 multi-head self-attention layers. As illustrated in Figure 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). The overall mapping function of the visual encoder can be summarized as follows: $\mathcal{F}_{\mathcal{E}_v} : V \to \mathbf{z}_v$.

4.3. Audio Encoder

The goal of the ViT-based [23] audio encoder $\mathcal{F}_{\mathcal{E}_a}$ is to learn meaningful features from the raw input audio

$A \supset \{A_i\}_{i=1}^n$. Following previous work [36, 87], 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 $\mathcal{F}_{\mathcal{E}_a}$[4]. 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. The overall mapping function of the audio encoder can be summarized as follows: $\mathcal{F}_{\mathcal{E}_a} : A' \to \mathbf{z}_a$.

4.4. Frame Classification Module

We further deploy frame-level classification modules on top of the audio and visual latent features. Let's denote the ground truth labels for visual and audio as $Y_v^{(f)}$ and $Y_a^{(f)}$, respectively. The visual classification module $\mathcal{F}_{\mathcal{C}_v}$ maps the latent visual features $\mathbf{z}_v$ to labels $\hat{Y}_v^{(f)} \in \mathbb{R}^T$ via $\mathcal{F}_{\mathcal{C}_v} : \mathbf{z}_v \to \hat{Y}_v^{(f)}$. Similarly, the audio classification module $\mathcal{F}_{\mathcal{C}_a}$ maps latent audio features $\mathbf{z}_a$ to labels $\hat{Y}_a^{(f)} \in \mathbb{R}^T$ via $\mathcal{F}_{\mathcal{C}_a} : \mathbf{z}_a \to \hat{Y}_a^{(f)}$.

4.5. Boundary Localization Module

This module facilitates the learning of deepfake localization. Motivated by BSN++ [78], we adopt the proposal relation block (PRB) as the framework for the boundary

---

[4]Note that we only incorporate the multi-head self-attention layers of the ViT for the audio encoder, i.e. the patching and positional encoding is purposefully omitted due to computational constraints of BA-TFD+ framework.

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;

     $\mathbf{z}_v \leftarrow \mathcal{F}_{\mathcal{E}_v}(V)$;

     $\mathbf{z}_a \leftarrow \mathcal{F}_{\mathcal{E}_a}(\text{mel-spetrogram}(A))$;

     $Y^{(c)} \leftarrow \eta_{vi} \wedge \eta_{ai}$;

     $\mathcal{L}_c \leftarrow \text{ContrastiveLoss}(\mathbf{z}_v, \mathbf{z}_a, Y^{(c)})$;

     $\hat{Y}_v^{(f)} \leftarrow \mathcal{F}_{\mathcal{C}_v}(\mathbf{z}_v)$;

     $\hat{Y}_a^{(f)} \leftarrow \mathcal{F}_{\mathcal{C}_a}(\mathbf{z}_a)$;

     /* 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;

---

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 j-th 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 $\hat{Y}_v^{(b)(pc)} \in \mathbb{R}^{D \times T}$. The overall mapping function can be summarized as follows: $\mathcal{F}_{\mathcal{B}_v} : \mathbf{z}_v \rightarrow (\hat{Y}_v^{(b)(p)}, \hat{Y}_v^{(b)(c)}, \hat{Y}_v^{(b)(pc)})$.

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 convolution layer. The overall mapping function can be summarized as follows: $\mathcal{F}_{\mathcal{B}_a} : \mathbf{z}_a \rightarrow (\hat{Y}_a^{(b)(p)}, \hat{Y}_a^{(b)(c)}, \hat{Y}_a^{(b)(pc)})$.

4.6. Multimodal Fusion Module

The fusion module illustrated in Figure 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 audio and visual 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 $W_a \in \mathbb{R}^{D \times T}$. Finally, we calculate element-wise weighted average for the fusion boundary map predictions $\hat{Y}^{(b)(p)}$, $\hat{Y}^{(b)(c)}$ and $\hat{Y}^{(b)(pc)}$. 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 proven to be helpful [17, 19] to eliminate the misalignment between different modalities (*i.e.* visual and audio). Motivated by this objective, BA-TFD+ uses the latent audio and visual features $\mathbf{z}_v$ and $\mathbf{z}_a$ of real videos acting 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 audio and visual modalities for the 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 fake or real.

4.7.3 Boundary Matching Loss

Following the standard protocol [52, 78], 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 value of possible 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.

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 ($\mathcal{L}_b$). The multimodal boundary matching loss is defined as follows,

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

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

where, m is the modality (visual v or audio a), η_m specifies whether modality m is modified, α is on of the boundary map types from the boundary module, $Y_\phi^{(b)} \in 0^{D \times T}$ 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 the three types of fusion boundary map outputs - 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 [78], we averaged the three boundary maps to produce the final boundary map output $\hat{Y}^{(b)}$. The boundary map represents the confidence for all proposals in the video. Since this operation produces duplicated proposals, we post-processed the proposals with Soft Non-Maximum Suppression (S-NMS) [6] similar to BSN++ [78].

5. Experiments

5.1. Data Partitioning

We split the LAV-DB 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 [52, 78], we additionally prepared a subset of the full set denoted as *subset* where the audio manipulated data is removed.

Table 3. **Temporal forgery localization results on the 'fullset' of the proposed dataset.** The visual-only version of the proposed method uses the output from the visual boundary matching layer, showing 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 [52]	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 [17]	12.78	01.62	00.00	37.88	36.71	34.39	32.15
AGT [61]	17.85	09.42	00.11	43.15	34.23	24.59	16.71
BSN++ [78]	56.41	32.57	00.21	74.93	71.11	64.98	59.29
AVFusion [3]	65.38	23.89	00.11	62.98	59.26	54.80	52.11
BA-TFD [11]	79.15	38.57	00.24	67.03	64.18	60.89	58.51
TadTR [57]	80.22	61.04	**05.22**	72.50	72.50	70.56	69.18
ActionFormer [90]	85.23	59.05	00.93	77.23	77.23	77.19	76.93
TriDet [74]	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 proposed dataset.** The visual-only version of the proposed method uses the output from the visual boundary matching layer, showing 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 [52]	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 [17]	23.43	03.48	00.00	58.53	56.68	53.16	49.67
AGT [61]	15.69	10.69	00.15	49.11	40.31	31.70	23.13
BSN++ [78]	65.26	37.70	00.22	78.89	76.32	71.00	65.38
AVFusion [3]	62.01	22.77	00.11	61.98	58.08	53.31	50.52
BA-TFD [11]	85.20	47.06	00.29	67.34	64.52	61.19	59.32
TadTR [57]	83.48	63.57	**05.44**	74.15	74.15	72.42	71.38
ActionFormer [90]	79.48	48.01	01.08	70.38	70.38	70.36	70.08
TriDet [74]	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

5.2. Implementation Details

The proposed framework BA-TFD+ is implemented in PyTorch [66] using 2 NVIDIA A100 80GB GPUs. We resized the input videos to 96×96 to reduce the computational cost of the MViTv2-based visual backbone. Similarly, the temporal dimension T is fixed to 512 for LAV-DF and 300 for ForgeryNet [33] and DFDC [22]. The latent space features $\mathbf{z}_a$ and $\mathbf{z}_v$ 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 process proposed in MViTv2 [50]. 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-DB dataset for deepfake detection and localization tasks.

- **Deepfake Detection.** Following standard evaluation protocols [22, 68], we use Area Under the Curve (AUC) as evaluation metric for binary deepfake classification task.

- **Deepfake Localization** We are the first to benchmark deepfake localization task. We adopt Average Precision (AP) and Average Recall (AR) as the evaluation metrics following prior literature [33]. For AP, we set the IoU thresholds to 0.5, 0.75 and 0.95, following ActivityNet's [9] evaluation protocol. Similarly for AR, as 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] respectively. When evaluat-

Table 5. **Temporal forgery localization on ForgeryNet dataset.** Performance comparison of BA-TFD+ for temporal forgery localization on ForgeryNet dataset.

Method	Avg. AP	AP@0.5	AP@0.75	AP@0.95	AR@5	AR@2
Xception [16]	62.83	68.29	62.84	58.30	73.95	25.83
X3D-M+BSN [24,54]	70.29	80.46	77.24	55.09	86.88	81.33
X3D-M+BMN [24,52]	83.47	90.65	88.12	74.95	91.99	88.44
SlowFast+BSN [25,54]	73.42	82.25	80.11	60.66	88.78	83.63
SlowFast+BMN [25,52]	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

Table 6. **Deepfake Detection.** Fake or Real classification performance comparison on the DFDC dataset.

Method	AUC
Meso4 [1]	0.753
FWA [49]	0.727
Siamese [59]	0.844
MDS [17]	0.916
BA-TFD [11]	0.846
BA-TFD+ (ours)	**0.937**

ing the proposed approach on ForgeryNet [33], we follow the protocol mentioned 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 [33], we used only the visual pipeline of BA-TFD+ 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 [22], 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 [52], AGT [61], AVFusion [3], MDS [17], BSN++ [78], TadTR [57], ActionFormer [90], and TriDet [74]. 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 [12]. For the methods that require S-NMS [6] 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 fullset and subset of the LAV-DF dataset for temporal forgery lo-

calization.

Full Set. From Table 3, our method achieves the best performance for AP@0.5 and AR@100. Unlike temporal action localization datasets, in our dataset, there is a single label for the fake segments. Thus, it is reasonable that the AP score is relatively high. The multimodal MDS method is not designed for temporal forgery localization tasks and predicts only fixed length segments (*i.e.* cannot predict the precise boundaries), hence the scores for that method 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 have relatively better performance as they are one-stage temporal action localization approaches that generate more precise segments. Additionally, we observe the BMN trained with an end-to-end visual encoder performs better than using pre-trained I3D features. With the multimodal complimentary information, our approach BA-TFD+ outperforms the aforementioned approaches.

Subset. 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 in [33]). As shown in Table 5, the performance of the visual-only BA-TFD+ exceeds the previous best model SlowFast [25]+BMN [52], 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 DFDC [22] following the

Table 7. **Ablation Studies (Loss functions).** The contribution of different losses in the BA-TFD+. Temporal forgery localization results are computed on the full set of the proposed 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. **Ablation Studies (Pre-trained Features).** Comparison of different pre-trained features in visual-only modality. E2E denotes end-to-end training.

Visual	Audio	Citation	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
I3D	E2E	[12]	74.76	59.57	04.02	74.28	71.92	68.64	66.63
MARLIN	E2E	[10]	92.27	75.11	04.10	77.93	76.38	74.53	73.47
3DMM	E2E	[5]	01.84	00.11	00.00	34.00	31.54	20.94	11.81
E2E	TRILLsson3	[75]	95.16	82.67	05.65	81.21	79.80	78.22	77.49
E2E	Wav2Vec2	[2]	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. **Ablation Studies.** Comparison of different backbone architectures for visual-only modality of BA-TFD+.

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**

configuration of [17]. As shown in Table 6, the performance of our method is better than previous methods such as Meso4 [1], FWA [49], Siamese [59], and MDS [17]. In summary, our method performs well on the classification task.

6.3. Ablation Studies

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 [91]. 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 model's 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 the boundary module mechanism have localization ability to detect the fake segments more precisely.

Impact of Pre-Trained Features. From the previous literature [57, 90], pre-trained visual features, such as I3D [12], are commonly used for temporal action localization methods. Since the I3D features are pre-trained on the Kinetics dataset [41], 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-
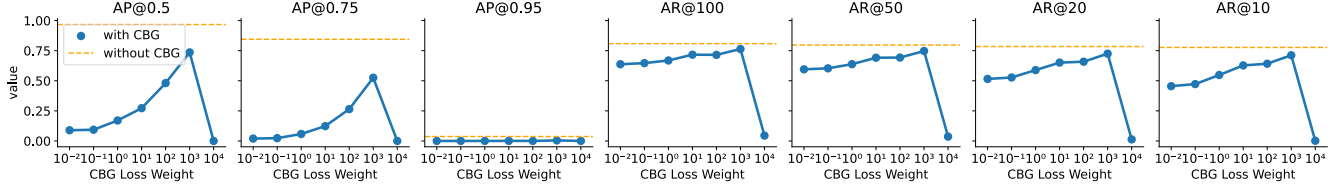
Figure 6. **Ablation Studies (CBG module).** The figure shows the comparison of CBG vs withour CBG prediction of the confidence score of the timestamp i.e. starting point or ending point of the segments in the temporal axis (See Section 6.3 for maore details)

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 [10] and 3DMM [5]) and audio features (TRILLsson [75] and Wav2Vec2 [2]). 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.

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 displayed in Table 9. Compared to the previous model BA-TFD [11] as baseline (3D-CNN + CNN + BMN [52]), we used the attention-based architectures including MViTv2 [50] and ViT [23] families for encoders and attention-based BSN++ modules [78] 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.

Impact of CBG in Boundary Matching Module We adopted the method from BSN++ [78] to improve the performance for temporal localization. This method includes two modules, complementary boundary generator (CBG) and proposal relation block (PRB). The CBG module pre-dicts the confidence of the timestamp is a starting point or an ending point of the segments in the temporal axis. The PRB module which is based on BMN [52] predicts the boundary map which contains the confidences of dense segment proposals. For inference, the results from both modules will be multiplied as the final output. In this ablation study, we aim to discuss the impact of the CBG module.

We trained several models containing the CBG module with different CBG loss weights from 10^{-2} to 10^4, and also a model without the CBG module. As shown in Figure 6, the best CBG loss weight among the models is 10^3. However, compared with the non-CBG model, the best model with CBG can only compete on AR metrics 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

This work introduces and investigates a novel problem related to content-driven deepfake generation and detection. To this end, we propose a new dataset in which the audio and visual are modified at specific locations based on the change in sentiment of the content. We also propose a new method for temporal forgery localization in such partially modified videos. The experiments conducted show that our method achieves better performance than previous relevant state-of-the-art methods.

Ethical Concerns. The proposed dataset potentially might have a negative social impact. Since the individuals in the dataset are celebrities, the content in the dataset may be used for unethical purposes such as spreading fake rumours. Also, the dataset generation pipeline can be used to create fake videos. To encounter the potential negative impact of our work, we prepared a license for public usage of the dataset and proposed the temporal forgery localization method.

Limitations. This work has some limitations: 1) The audio reenactment method used in the dataset does not always generate the reference style, 2) The resolution of the dataset is constrained on the basis of source videos and 3) The high score of the classification results indicates the necessity of improving the visual reenactment method.

Future Work. Major improvement in the future will be extending the generation pipeline to include word tokens in-

sertion, substitution and deletion and converting statements into questions.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, Dec. 2018. ISSN: 2157-4774.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[3] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla. Hear Me Out: Fusional Approaches for Audio Augmented Temporal Action Localization, Aug. 2021. arXiv: 2106.14118 version: 3.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., June 2009. Google-Books-ID: KGIbfiiP1i4C.

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99*, pages 187–194, Not Known, 1999. ACM Press.

[6] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS – Improving Object Detection With One Line of Code. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5561–5569, 2017.

[7] John Brandon. There Are Now 15,000 Deepfake Videos on Social Media. Yes, You Should Worry. *Forbes*, Oct. 2019.

[8] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. *Procdings of the British Machine Vision Conference 2017*, May 2019. Publisher: British Machine Vision Association.

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[10] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. MARLIN: Masked Autoencoder for facial video Representation LearnINg, Nov. 2022. arXiv:2211.06627 [cs].

[11] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10, Nov. 2022.

[12] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[13] Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti. SC-GlowTTS: an Efficient Zero-Shot Multi-Speaker Text-To-Speech Model, June 2021. arXiv: 2104.05557.

[14] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-Head Generation with Rhythmic Head Motion. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 35–51, Cham, 2020. Springer International Publishing.

[15] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019.

[16] Francois Chollet. Xception: Deep Learning With Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[17] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. 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, pages 439–447, New York, NY, USA, Oct. 2020. Association for Computing Machinery.

[18] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*, 2018.

[19] Joon Son Chung and Andrew Zisserman. Out of Time: Automated Lip Sync in the Wild. In Chu-Song Chen,

Jiwen Lu, and Kai-Kuang Ma, editors, *Computer Vision – ACCV 2016 Workshops*, Lecture Notes in Computer Science, pages 251–263, Cham, 2017. Springer International Publishing.

[20] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining EfficientNet and Vision Transformers for Video Deepfake Detection. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, Lecture Notes in Computer Science, pages 219–229, Cham, 2022. Springer International Publishing.

[21] Oscar de Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George. Deepfake Detection using Spatiotemporal Convolutional Networks, June 2020. arXiv: 2006.14749.

[22] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset, Oct. 2020. arXiv: 2006.07397.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[24] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[25] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

[26] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998. Google-Books-ID: Rehu8OOzMIMC.

[27] Jiyang Gao, Kan Chen, and Ram Nevatia. CTAP: Complementary Temporal Action Proposal Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–83, 2018.

[28] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3628–3636, 2017.

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, Oct. 2020.

[30] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal Inconsistency Learning for DeepFake Video Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3473–3481. Association for Computing Machinery, New York, NY, USA, Oct. 2021.

[31] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. DeepFake Detection by Analyzing Convolutional Traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 666–667, 2020.

[32] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.

[33] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4360–4369, 2021.

[34] Young-Jin Heo, Woon-Ha Yeo, and Byung-Gyu Kim. DeepFake detection algorithm based on improved vision transformer. *Applied Intelligence*, 53(7):7512–7527, Apr. 2023.

[35] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS Challenge on Action Recognition for Videos "in the Wild". *Computer Vision and Image Understanding*, 155:1–23, Feb. 2017. arXiv: 1604.06182.

[36] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, 136:110124, Mar. 2023.

[37] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You Said That?: Synthesising Talking Faces from Audio. *International Journal of Computer Vision*, 127(11):1767–1779, Dec. 2019.

[38] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multi-speaker text-to-speech synthesis. In *Proceedings of*

the 32nd International Conference on Neural Information Processing Systems, NIPS'18, pages 4485–4495, Red Hook, NY, USA, Dec. 2018. Curran Associates Inc.

[39] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020.

[40] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 1428–1436, New York, NY, USA, Oct. 2019. Association for Computing Machinery.

[41] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset, May 2017. arXiv:1705.06950 [cs].

[42] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S. Woo. Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors. *Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection*, pages 7–15, Oct. 2021. arXiv: 2109.02993.

[43] Hasam Khalid, Shahroz Tariq, and Simon S. Woo. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset, Aug. 2021. arXiv: 2108.05080.

[44] Davis E. King. Dlib-ml: A Machine Learning Toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, Dec. 2009.

[45] Pavel Korshunov and Sebastien Marcel. DeepFakes: a New Threat to Face Recognition? Assessment and Detection, Dec. 2018. arXiv: 1812.08685.

[46] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast Face-Swap Using Convolutional Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017.

[47] Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. OpenForensics: Large-Scale Challenging Dataset for Multi-Face Forgery Detection and Segmentation In-the-Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10117–10127, 2021.

[48] John K. Lewis, Imad Eddine Toubal, Helen Chen, Vishal Sandesera, Michael Lomnitz, Zigfried Hampel-Arias, Calyam Prasad, and Kannappan Palaniappan. Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9, Mar. 2020. ISSN: 2332-5615.

[49] Yuezun Li and Siwei Lyu. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 7, 2019.

[50] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[51] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[52] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2019.

[53] Tianwei Lin, Xu Zhao, and Zheng Shou. Single Shot Temporal Action Detection. In *Proceedings of the 25th ACM international conference on Multimedia*, MM '17, pages 988–996, New York, NY, USA, Oct. 2017. Association for Computing Machinery.

[54] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[55] Xiaolong Liu, Song Bai, and Xiang Bai. An Empirical Study of End-to-End Temporal Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20010–20019, 2022.

[56] Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip H. S. Torr. Multi-Shot Temporal Event Localization: A Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12596–12606, 2021.

[57] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-End

Temporal Action Detection With Transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022. Conference Name: IEEE Transactions on Image Processing.

[58] Yisroel Mirsky and Wenke Lee. The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, 54(1):7:1–7:41, Jan. 2021.

[59] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 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, pages 2823–2832, New York, NY, USA, Oct. 2020. Association for Computing Machinery.

[60] Daniel Mas Montserrat, Hanxiang Hao, Sri K. Yarlagadda, Sriram Baireddy, Ruiting Shao, Janos Horvath, Emily Bartusiak, Justin Yang, David Guera, Fengqing Zhu, and Edward J. Delp. Deepfakes Detection With Automatic Face Weighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 668–669, 2020.

[61] Megha Nawhal and Greg Mori. Activity Graph Transformer for Temporal Action Localization. *arXiv:2101.08540 [cs]*, Jan. 2021. arXiv: 2101.08540.

[62] Paarth Neekhara, Shehzeen Hussain, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. Expressive Neural Voice Cloning. In *Proceedings of The 13th Asian Conference on Machine Learning*, pages 252–267. PMLR, Nov. 2021. ISSN: 2640-3498.

[63] Dufou Nick and Jigsaw Andrew. Contributing Data to Deepfake Detection Research, Sept. 2019.

[64] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019.

[65] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, Sept. 2016. arXiv: 1609.03499.

[66] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[67] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 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, pages 484–492, New York, NY, USA, Oct. 2020. Association for Computing Machinery.

[68] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[69] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning Internal Representations by Error Propagation. Technical report, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE, Sept. 1985. Section: Technical Reports.

[70] Ian Sample. What are deepfakes – and how can you spot them? *The Guardian*, Jan. 2020.

[71] Conrad Sanderson, editor. *The VidTIMIT Database*. IDIAP, 2002.

[72] Oscar Schwartz. You thought fake news was bad? Deep fakes are where truth goes to die. *The Guardian*, Nov. 2018.

[73] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, Apr. 2018. ISSN: 2379-190X.

[74] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. TriDet: Temporal Action Detection with Relative Boundary Modeling, Mar. 2023. arXiv:2303.07347 [cs].

[75] Joel Shor and Subhashini Venugopalan. TRILLsson: Distilled Universal Paralinguistic Speech Representations. In *Interspeech 2022*, pages 356–360, Sept. 2022. arXiv:2203.00236 [cs, eess].

[76] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. 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*, pages 5734–5743, 2017.

[77] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal Action Localization in Untrimmed Videos via Multi-Stage CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016.

[78] Haisheng Su, Weihao Gan, Wei Wu, Yu Qiao, and Junjie Yan. BSN++: Complementary Boundary Regressor with Scale-Balanced Relation Modeling for Temporal Action Proposal Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2602–2610, May 2021. Number: 3.

[79] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural Voice Puppetry: Audio-Driven Facial Reenactment. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV 2020*, Lecture Notes in Computer Science, pages 716–731, Cham, 2020. Springer International Publishing.

[80] Daniel Thomas. Deepfakes: A threat to democracy or just a bit of fun? *BBC News*, Jan. 2020.

[81] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2018.

[82] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, ICMR '22, pages 615–623, New York, NY, USA, June 2022. Association for Computing Machinery.

[83] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis, Apr. 2017. arXiv: 1703.10135.

[84] Deressa Wodajo and Solomon Atnafu. Deepfake Video Detection Using Convolutional Vision Transformer, Mar. 2021. arXiv: 2102.11126.

[85] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-TAD: Sub-Graph Localization for Temporal Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020.

[86] Ke Yang, Peng Qiao, Dongsheng Li, Shaohe Lv, and Yong Dou. Exploring Temporal Preservation Networks for Precise Temporal Action Localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. Number: 1.

[87] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023. Conference Name: IEEE Transactions on Information Forensics and Security.

[88] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, May 2019. ISSN: 2379-190X.

[89] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph Convolutional Networks for Temporal Action Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7094–7103, 2019.

[90] Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing Moments of Actions with Transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 492–510, Cham, 2022. Springer Nature Switzerland.

[91] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal Action Detection With Structured Segment Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.

[92] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021.

[93] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. MakeltTalk: speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6):221:1–221:15, Nov. 2020.

[94] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pages 2382–2390, New York, NY, USA, Oct. 2020. Association for Computing Machinery.