Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization

Supplementary Material

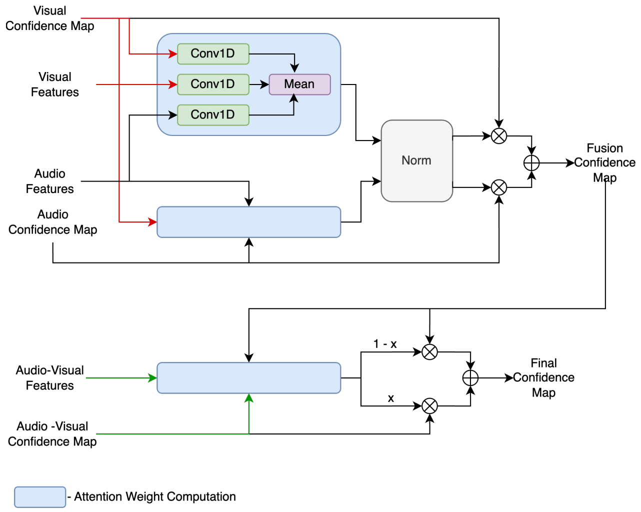

Figure 1. **Fusion module:** The fusion module integrates visual, audio, and audio-visual features using attention-weighted confidence maps to generate a refined final confidence map. This process ensures effective multimodal fusion by balancing contributions from individual and combined modalities.

1. Fusion Module for Localization

The localization module, including the boundary and fusion modules, is adapted from BA-TFD [5]. In the original design, the fusion module processes visual and audio confidence maps and features to generate the final confidence map using attention weight computation. We enhanced this by incorporating visual, audio, and audio-visual confidence maps and features into the fusion process, while maintaining a similar strategy. As illustrated in Fig. 1, the fusion module computes a $(D \times T)$-dimensional attention weight for each modality using its features, confidence map, and the other modality's features. These attention weights are normalized to produce modality-specific weights, which are then multiplied by their respective confidence maps and added to form the fusion confidence map. Subsequently, the audio-visual features, confidence maps, and fusion confidence map undergo another round of attention-weight computation. The output is split into two branches: one is multiplied by the audio-visual confidence map, while the other undergoes a complement operation (1 minus its value) before being multiplied by the fusion confidence map. Finally, the results from both branches are combined to produce the final output confidence map.

2. Implementation Details

The videos are extracted at a rate of 25fps and audio at 16kHz. The audio is converted to a mel spectrogram created with 64 mel filterbanks, a window size of 321, and a hop length of 160.

To keep the encoder models simple and light-weight, we used AV-HuBERT's [22] modified ResNet-18 visual encoder with 512 channels and ViT [9] with 192 hidden size, 768 MLP size, 12 layers, and 3 heads as audio encoder. The outputs of each of the encoders are then projected to a 256-dimensional feature, to match the feature dimension of visual and audio features (hence, f=256). Following Feng *et al*. [10], we selected the maximum offset for a temporal shift, τ to be 15, thus generating 31 scores for each frame. The segment size is chosen to equal τ, i.e., 15. The decoder variant of the synchronization model involves 3 layers of transformer decoder with 4 attention heads and 256 channels, while the sparse encoder variant includes 3 layers of sparse self-attention with 512 channels. In the case of the sparse encoder synchronization model, the 512-dimensional inferences are first projected to 256-dimensional features before passing to the classification or localization model.

The unimodal transformer encoders in the localization module have a depth of 2, 4 attention heads, and 256 channels. The number of attention blocks, L, in the classification module, is 3, while that in the localization module is 6. For synchronization pretraining, we generate Gaussian targets using a variance of 1.5. We experimented with variances of 1.0, 1.5, and 2.0, and found that a variance of 1.5 yielded the best performance.

3. Dataset Details

LRS2 [1]: Lip Reading Sentences 2 (LRS2) is a large-scale dataset from Oxford-BBc for audiovisual speech recognition. It consists of videos consisting of spoken sentences along with their face tracks. Their pretraining split consists of around 97k utterances. Following [10], we utilized one-third portion for synchronization pretraining.

VoxCeleb2 [7]: VoxCeleb2 comprises real YouTube videos featuring over 6k celebrities, including more than 1 million utterances in the development set and around 36k in the test set. Each video contains celebrity interviews along with their speech audio. The dataset is fairly gender-balanced, with 61% male speakers, and represents diverse ethnicities, accents, professions, and age groups. It includes videos captured in various challenging visual and auditory conditions. Following the LRS2 setup, we pretrain our synchronization model on a 32k samples from this dataset.

FakeAVCeleb [15]: The FakeAVCeleb dataset is designed for deepfake detection and includes a total of 20,000 video clips. It features 500 real videos sourced from VoxCeleb2 [7] and 19,500 deepfake samples generated using various manipulation techniques, such as Faceswap [16], FaceswapGAN [19], Wav2Lip [21], and SV2TTS [13]. The dataset is divided into four categories: Real Visuals - Real Audio, Fake Visuals - Real Audio, Fake Visuals - Fake Audio, and Real Visuals - Fake Audio.

KoDF [17]: KoDF is a large-scale video deepfake dataset comprising of 403 Korean subjects. The dataset consists of 62,000+ real videos and 175,000+ deepfakes generated using 6 different synthesis methodologies, aimed at a better generalization to real-world scenarios.

DFDC [8]: The DFDC dataset is a large multimodal deepfake dataset containing over 100,000 samples from more than 3,400 subjects. It includes videos generated using seven visual manipulation techniques and one audio swap method. The dataset features videos recorded under challenging lighting conditions, varied human poses, and diverse camera angles. Following previous works [11, 12, 20], we filtered the videos to include only those with a single person, where facial landmarks were successfully detected, and sampled 3,215 videos for testing.

LavDF [3, 5]: The LAV-DF dataset features deepfakes where only specific segments of a video are manipulated. Similar to FakeAVCeleb, it comprises around 36,000 real samples from VoxCeleb2, with manipulations applied to either or both modalities, resulting in over 99,000 deepfake samples. The dataset includes 114,253 fake segments ranging in duration from 0 to 1.6 seconds, with an average length of 0.65 seconds. Notably, 89.26% of fake segments are shorter than 1 second. Videos in the dataset have a maximum length of 20 seconds, with 69.61% being under 10 seconds. Modality modifications are evenly distributed among four types: visual-modified, audio-modified, both-modified, and real.

AVDF1M [4]: The AV-Deepfake 1M (AVDF1M) dataset, a successor to LAV-DF, contains deepfakes with manipulations in small segments. It is significantly larger than LAV-DF, comprising over 1 million samples from more than 2,000 subjects. Compared to LAV-DF, AVDF1M includes a wider range of deepfake segment lengths, with an average manipulated segment length of 0.326 seconds. For training, we sampled 74,614 and 7,387 videos from the dataset's training set for our training and validation subsets, respectively. Model evaluation was performed on the original validation set, which contains 57,340 samples.

Dataset	AP	AUC
DFDC [8]	98.4	85.2
CREMA [24]	99.8	99.7

Table 1. **Cross-dataset Performance:** We report the classification performance achieved by the 'Ours (Decoder, VoxCeleb2)' model when tested on other deepfakes.

4. Cross-Dataset Generalization on other Deepfakes

To assess the cross-dataset generalizability of our model further, we test our 'Ours (Decoder, VoxCeleb2)' model, trained on FakeAVCeleb [15] on the following datasets, and report the result in Tab. 1:

- **DFDC [8]:** We use a subset of samples from the dataset following the evaluation protocol of [11, 12, 20]. The model achieved an AUC of 85.15% and an AP of 98.44%, demonstrating strong performance even on videos captured under challenging lighting and camera conditions.
- **CREMA [24]:** Following recent use of diffusion models for talking head generation, we adopt the setup similar to [14] and use the generation outputs from [24], which include 820 test samples. We augment this with 820 real videos from VoxCeleb2[7], achieving an AP of 99.8 and an AUC of 99.7, indicating strong performance on diffusion-generated talking head videos.

5. Additional Ablation Results

5.1. The special case of FVRA

While our model can perform efficiently under cross-manipulation generalization for the categories mentioned in Tab. 4, we evaluated the 'Ours (Decoder, VoxCeleb2)' model under 3 additional categories namely FVRA-WL(real audio with fake video by Wav2Lip), FVRA-FS(real audio with fake video by Faceswap) and FVRA-GAN(real audio with fake video by FaceswapGAN), and observed that the AUC values achieved by the model under cross-manipulation generalization for these 3 categories collapsed to near-chance values. This can be explained by the model's strong reliance on the audio modality, which is discussed in detail in Sec. 5.2.2. However, for cross-dataset generalization, the deepfake samples in KoDF [17] are generated using audio-driven methods, ATFHP [25] and Wav2Lip [21], which utilize real audio to create deepfake videos. This setup aligns with the FVRA category but draws from an unseen distribution. As shown in Tab. 5, our model effectively identifies deepfakes within these test samples.

Feature Set	AP	AUC
Synchronization Score	98.9	68.6
Submodule Features, Shift=0	99.6	85.5
Ours (Submodule Features, Shift=-τ)	**99.7**	**88.0**

Table 2. **Synchornization Score vs Submodule Features:** We report the classification performance achieved by the model when trained on different sets of features predicted by the pretraining model.

5.2. Feature Set Analysis

5.2.1. Synchornization Score vs Submodule Features

We tried to evaluate the impact of different feature combinations, inferred by our pretrained model, over the classification task performance. To test this, we first compared three possible feature sets. Firstly we directly used the final prediction of our pretraining model, which are $3\ T \times 31$ dimensional features that denote the frame-level cross-modal (Γ_{V-A}) and intra-modal (Γ_{V-V} and Γ_{A-A}) synchronization scores. Since the classification model processes $T \times f$ dimensional input data, each 31 frame-level synchronization score is projected into a f-dimensional space.

Secondly, instead of the final synchronization score, we decided to investigate whether features predicted by submodules inside the pertaining model could hold enough information about the cross-modal and intra-modal consistency. Hence, for cross-modal, we picked up the features by the attention module first, without any shift (Shift=0), and second, with maximum shift (Shift=τ) to the audio encoding. For the decoder variant, we selected the features after passing through every decoder layer (before computing frame level similarity) and similarly, for the sparse encoder variant, we selected the features after passing through the sparse attention block (before passing to the final feedforward block). For intra-modal consistency, we directly picked up the features predicted by the individual encoders. Although each of the features is $T \times f$ dimensional, to maintain architectural consistency for the classification task each of the features is projected to a f-dimensional space.

Tab. 2 shows the performance of each of the three feature sets, inferred by the 'Ours (Decoder, VoxCeleb2)' model and used to train our classification model. The results show that the submodule features have a positive impact on the classification task. This shows that the individual submodules of the pertaining model can capture information relevant to cross-modal and intra-modal synchronization. Additionally, we feel that the relatively poor performance produced by using the final synchronization score can be attributed to the low dimensionality of the feature, which may not be able to capture all the information required to train the classification model of similar size. In other words, direct synchronization scores are not enough and are not as impactful as the high-dimensional features holding the synchronization information for deepfake classification. Additionally, submodule features with no shift and maximum shift performed similarly, but due to a marginal difference, we selected the maximum shift features as our final feature.

5.2.2. Contribution of Submodule Features

To understand the impact of each feature used by the classification head, we trained the model using only one feature at a time instead of all three. Since this setup involves a single feature, alternating cross-attention is unnecessary. Therefore, we replaced the transformer decoder layers in the classification module with an equal number of transformer encoder layers. Tab. 3 reports the results for the three individual feature sets: 'Only A-V Submodule Features,' 'Only Video Embeddings,' and 'Only Audio Embeddings.' A significant drop in performance is observed compared to our proposed model, which utilizes all three features. This suggests that neither cross-modal nor uni-modal features alone are sufficient. Notably, the minimal performance drop when using only audio embeddings indicates the model's strong reliance on the audio modality. This reliance could explain the poor performance in the Fake Visuals–Real Audio (FVRA) category in FakeAVCeleb.

To evaluate the impact of using uni-modal embeddings instead of intra-modal synchronization model outputs, we conducted two experiments. First, we trained the classification module with 'Only A-A Submodule Features' and 'Only V-V Submodule Features.' The results show a slight performance drop compared to their respective uni-modal embeddings. Next, we replaced the uni-modal embeddings in our original classification module with the corresponding synchronization submodule features. We observed that uni-modal embeddings slightly outperformed the submodule features, with a 0.1% increase in AP and a 3.8% gain in AUC. We conjecture that since the video and audio encoders were shared across all cross-modal and intra-modal synchronization losses, the unimodal embeddings effectively capture the necessary and sufficient information for intra-modal temporal synchronization to be used in the second stage. Additionally, replacing video and audio embeddings with A-A and V-V submodule features increases computational complexity. The inference time of the pretrained synchronization model rises from 20.60 ms to 28.37 ms, and the FLOPs count increases from 324.58 to 392.32 GFLOPs.

It is also worth noticing that while the impact in AP is not much, AUC is significantly impacted in the experiments. This could be explained by the imbalance in the FakeAVCeleb [15] with only 500 real videos but 19,500 deepfakes in the entire dataset.

Feature Set Type	Feature set	AP	AUC
Submodule Features	A-V + A-A + V-V	99.6	84.2
	Only A-V	98.4	53.8
	Only V-V	97.4	37.7
	Only A-A	99.1	70.2
Embeddings	Only Video	97.5	39.6
	Only Audio	99.5	81.9
Ours	(A-V Submodule features + Video Embeddings + Audio Embeddings)	**99.7**	**88.0**

Table 3. **Contribution of Submodule Features on Performance:** We report the classification performance achieved by the model when trained on individual embeddings and synchronization submodule features. The best results are highlighted in bold

Method	RVFA		FVFA-FS		FVFA-GAN		FVFA-WL		AVG-FV	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
Ours (Decoder, VoxCeleb2)	**98.6**	**97.9**	**99.6**	**98.1**	**99.7**	**97.4**	**99.6**	**96.7**	**99.6**	**97.4**
Ours (Decoder, LRS2)	96.7	94.6	99.3	96.3	99.5	96.1	99.5	95.3	99.4	95.9
Ours (Sparse Encoder, LRS2)	94.0	95.4	99.3	96.6	99.3	95.2	99.0	90.9	99.2	94.2
Ours (Sparse Encoder, VoxCeleb2)	96.5	95.8	98.3	91.2	99.2	93.9	99.4	94.1	98.9	93.1

Table 4. **Cross-Manipulation Detection on FakeAVCeleb:** We report the Average Precision(AP) and AUC scores over one manipulation category while training on the rest of the data. For this, we consider four categories: (i) **RVFA:** Real Visual - Fake Audio (SV2TTS [13]),(ii) **FVFA-FS:** Fake Visual - Fake Audio (FaceSwap [16] + Wav2Lip [21] + SV2TTS [13]), (iii) **FVFA-GAN:** Fake Visual - Fake Audio (FaceSwapGAN [19] + Wav2Lip [21] + SV2TTS [13]), and (iv) **FVFA-WL:** Fake Visual - Fake Audio (Wav2Lip [21] + SV2TTS [13]). The column **AVG-FV** refers to the mean of the performance achieved on the four Fake Visual categories. Bold highlights the best performance for every metric under each category.

Method	KoDF [17]	
	AP	AUC
Ours (Decoder, VoxCeleb2)	**98.9**	99.0
Ours (Decoder, LRS2)	56.7	70.1
Ours (Sparse Encoder, LRS2)	98.6	**99.2**
Ours (Sparse Encoder, VoxCeleb2)	94.4	96.9

Table 5. **Cross-Dataset performance on KoDF:** AP and AUC score(%) achieved on KoDF dataset by the models trained on FakeAVCeleb. The best results are highlighted in bold.

5.3. Different Pretraining dataset

To understand whether the choice of the dataset has any impact on the model's performance, we pretrained our two variants on two datasets, VoxCeleb2 and LRS2. And used the features inferred by them to evaluate the classification task. Tab. 4, and Tab. 5 report the performance of the two experiments under cross-manipulation and cross-dataset generalization respectively. It can be seen that the variants pretrained on VoxCeleb2 outperform the decoder variant pretrained on LRS2 under the majority of categories. This can be attributed to the fact that the FakeAVCeleb dataset consists of real videos sampled from the VoxCeleb2 dataset, causing both stages to witness data from the same distribution. Nevertheless, the performance achieved by the variants pretrained on LRS2 is on par with the one pretrained on VoxCeleb2 in each of the two settings. While the performance of the decoder variant pretrained on LRS2 and evaluated under the cross-dataset generalization is interesting and may need further study, the overall result by the sparse encoder shows that the model is not overfitting to specific data distribution and the synchronization features learned during the pretraining are stable across datasets.

5.4. Computational Analysis

To analyze the computational complexity of our model, we measured the inference time and FLOPs separately for inference, classification, and localization, as shown in Tab. 6. Each experiment was conducted with a batch size of 1 on an NVIDIA H100 80GB HBM3 GPU. The results indicate that the GFLOPs for the decoder and sparse encoder variants during inference are quite similar, though the sparse encoder has a slightly faster inference time. Nevertheless, training the sparse encoder variant is significantly faster than the decoder variant. Additionally, the classification

Stage	GFLOPs	Time (ms)
Decoder variant inference	324.58	20.60
Sparse Encoder variant inference	324.60	17.24
Classification	3.11	4.26
Localization	296.42	72.20

Table 6. **Computational analysis:** We report the inference time and computational complexity in terms of GFLOPs for individual stages of the proposed model.

Method	AP@0.5	AP@0.75	AP@0.95
MDS [6]	12.78	1.62	0.00
BMN [18]	24.01	7.61	0.07
AVFusion [2]	65.38	23.89	0.11
BA-TFD [3]	76.90	38.50	0.25
ActionFormer [26]	85.23	59.05	00.93
TriDet [23]	86.33	70.23	03.05
BA-TFD+ [5]	96.30	84.96	04.44
UMMAFormer [27]	**98.83**	**95.54**	**37.61**
(Decoder-Frozen, VoxCeleb2)	87.40	66.80	05.72
(Decoder-Finetune, VoxCeleb2)	97.79	86.34	08.94

Table 7. **Effect of fine-tuning Decoder variant on Temporal Localization on Lav-DF:** We report Average Precision(%) achieved at IoU thresholds of 0.5, 0.75, and 0.95 on the LAV-DF dataset. The best scores are highlighted in bold.

Method	AP@0.5	AP@0.75	AP@0.95
(Decoder-Frozen, VoxCeleb2)	23.43	3.48	0.00
(Decoder-Finetune, VoxCeleb2)	95.13	66.98	0.32

Table 8. **Temporal Localization on AVDF1M:** We report Average Precision(%) achieved at IoU thresholds of 0.5, 0.75, and 0.95 on the AVDF1M dataset.

task runs efficiently, with minimal execution time. Finally, it is important to note that the reported inference time for the localization task includes non-maximum suppression.

5.5. Audio Modality Missing

To evaluate the classification model's performance without audio input, we replaced the audio with a zero tensor and passed it through the synchronization and classification models. During testing, videos with only deepfaked audio were treated as real, while those with manipulated visuals were considered fake. The model achieved an Average Precision of 98.1% and an AUC score of 71.6%. Although the Average Precision dropped by only 1.6% compared to the original score with both modalities, the AUC score saw a significant decline of 16.4%. Nevertheless, considering the model was trained without handling missing modalities, its performance remains strong and well above chance.

5.6. Effect of Fine Tuning Synchronization Model on Localization

To assess the impact of a larger model on localization performance, we fine-tuned the synchronization model along with the localization head. Tab. 7 reports the average precision at IoU thresholds of 0.5, 0.75, and 0.95. '(Decoder-Frozen, VoxCeleb2)' refers to our model with a frozen decoder, while '(Decoder-Finetune, VoxCeleb2)' denotes the fine-tuned version. Fine-tuning significantly improved performance by 10.39% at 0.5 IoU, 19.54% at 0.75 IoU, and 3.22% at 0.95 IoU. Moreover, the model outperformed BA-TFD+ [5] and achieved performance comparable to the state-of-the-art UMMAFormer [27] at 0.5 IoU. However, fine-tuning increased the number of trainable parameters from 18.8M to 50.3M, and the FLOPs for localization rose from 296.42 GFLOPs to 622.02 GFLOPs. Additionally, this fine-tuning affected the model's ability to generalize across both classification and localization tasks.

6. Localization Performance on AVDF1M

We further evaluated our model's performance on the more challenging AVDF1M dataset [4] for the task of temporal deepfake localization. We tested with two model variations, '(Decoder-Frozen, VoxCeleb2)' and '(Decoder-Finetune, VoxCeleb2)', following the setup described in Sec. 5.6, with results presented in Tab. 8. The '(Decoder-Frozen, VoxCeleb2)' variant shows a notable performance drop, indicating that AVDF1M is significantly more challenging than LAV-DF. In contrast, the '(Decoder-Finetune, VoxCeleb2)' variant achieves strong results at IoU thresholds of 0.5 and 0.75, demonstrating the effectiveness of end-to-end finetuning on this challenging dataset.

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. In *arXiv:1809.02108*, 2018. 1

[2] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla. Hear me out: Fusional approaches for audio augmented temporal action localization. *arXiv preprint arXiv:2106.14118*, 2021. 5

[3] 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. IEEE, 2022. 2, 5

[4] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. *arXiv preprint arXiv:2311.15308*, 2023. 2, 5

[5] Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. Glitch in the matrix: A large scale benchmark for content driven audio-visual forgery detection and localization. *Computer Vision and Image Understanding*, 236:103818, 2023. 1, 2, 5

[6] 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*, pages 439–447, 2020. 5

[7] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 1, 2

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020. 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023. 1

[11] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 2

[12] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14950–14962, 2022. 2

[13] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018. 2, 4

[14] Sarthak Kamat, Shruti Agarwal, Trevor Darrell, and Anna Rohrbach. Revisiting generalizability in deepfake detection: Improving metrics and stabilizing transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 426–435, 2023. 2

[15] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021. 2, 3

[16] 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. 2, 4

[17] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10744–10753, 2021. 2, 4

[18] 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. 5

[19] 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. 2, 4

[20] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024. 2

[21] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, pages 484–492, 2020. 2, 4

[22] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 1

[23] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 5

[24] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024. 2

[25] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2

[26] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 5

[27] Rui Zhang, Hongxia Wang, Mingshan Du, Hanqing Liu, Yang Zhou, and Qiang Zeng. Ummaformer: A universal multimodal-adaptive transformer framework for temporal forgery localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8749–8759, 2023. 5