EMO-BOOST: Emotion-Augmented Audio-Visual Features for Improved Generalization in Deepfake Detection

Supplementary Material

We structure the supplementary materials as follows: first, Section 7 provides details on our employed datasets, FakeAVCeleb and DeepSpeak v2, used for training, validation, and testing our methods across different evaluation scenarios and for benchmarking against state-of-the-art methods. Further, in Section 8, we provide the detailed tabular representations of the comparison of our framework in the leave-one-out evaluation scenario on both datasets, along with the tabular comparison of different fusion methods implemented in EmoForensics in the in-domain evaluation on FakeAVCeleb and DeepSpeak v2, and stability in performance between EmoForensics and SIMBA on the leave-one-out evaluation setup on FakeAVCeleb.

7. Details on Dataset

In this section, we provide some details on how we split and utilise the FakeAVCeleb [18] and DeepSpeak v2 [2] datasets for benchmarking our proposed framework Emo-Boost. We create a validation split for both datasets, which is used for learning rate scheduling and early stopping. Furthermore, we also validate our design choices for our methods based on the performance on this split. Method and Family splits are constructed following the setup provided by [19]. As described in Section 4, we introduce a new *val-test* split for the purpose of validating our design choices in the cross-manipulation evaluation scenario. For every Method and Family split, the val-test split is constructed by uniformly sampling 20% of both real and fake videos from the corresponding test set. We use this split only for the purpose of model selection and hyperparameter tuning. The model does not observe this split during training, and only observes the in-domain train and validation split, the latter being used for learning rate scheduling and early stopping. For reporting our results, we use the entire test split as provided, ie, by adding the newly created val-test split back to the test split.

7.1. FakeAVCeleb

We construct the train, validation, and test splits in FakeAVCeleb following the same mechanism as in [19] in the ratio of 60%, 10%, and 30%, respectively, on the basis of provided identity annotations. Similarly, following [19], we use four Method splits (Faceswap, FSGAN, Wav2Lip, and RealTime Voice Cloning) and two Family splits (Face Animation and Lip Synthesis) for our cross-manipulation evaluation setup. In these Method and Family splits, we introduce the val-test split as described before. For example, for the Faceswap method split, while the train and val splits do not contain any fake video manipulated using the Faceswap method, the test set contains 929 such videos, along with 150 real videos. To construct our val-test split, we sample 20% of each class. Therefore, the final val-test split for leave-one-out evaluation of Faceswap contains 30 real videos and 186 fake videos. These videos are then added back to the test set when obtaining the final results. The train, val, and test splits for the standard in-domain evaluation setup contain **12,935**, **2,176**, and **6,455** samples, respectively.

7.2. DeepSpeak v2

For the construction of the validation split in DeepSpeak v2, we use 20% of the training data while the test set remains untouched. The Method and Family splits, including the internal val-test split, are constructed in the same manner as in FakeAVCeleb, following [19]. In DeepSpeak v2, we use 15 Method Splits and 3 Family Splits for the leave-one-out evaluation setup. The train, validation, and test splits of the standard in-domain evaluation setup for DeepSpeak v2 comprise **10,645**, **2,661**, and **3,279** samples, respectively.

8. Further Results

8.1. Leave-one-out Evaluation

As described in Section 5.2, we observe motivating results from our proposed framework in the leave-one-out evaluation setup on FakeAVCeleb. We present the detailed performance and its comparison with other state-of-the-art multimodal deepfake detectors in Table 6. **Emo-Boost**ed SIMBA achieves the highest AUC across all splits except for the Face Animation Family, where it is the second best, and also attains the best average performance.

Similarly, we provide the detailed comparison of Emo-Boosted SIMBA against the other deepfake detectors in the same evaluation setup on DeepSpeak v2 in Table 7. While **Emo-Boost**ed SIMBA shows minor improvements on splits such as *LivePortrait Real*, *LatentSync ElevenLabs*, and *LatentSync Speechify*, its average performance remains slightly below that of SIMBA.

8.2. Further Ablations

EmoForensics. We provide the detailed comparison of different fusion strategies used in the EmoForensics framework to obtain the joint representation on the in-domain

evaluation setup on both FakeAVCeleb and DeepSpeak v2 in Table 5.

Table 5. **Fusion Strategy Ablation.** Comparison of different fusion strategies within the EmoForensics framework on the in-domain evaluation setting for FakeAVCeleb and DeepSpeak v2, reported in terms of AUC score. Element-wise addition yields the best performance across both datasets. Best results are highlighted in **bold**, and second-best are underlined.

EmoForensics	FakeAVCeleb	DeepSpeak v2
w/ Element-wise Addition	**82.10**	**65.38**
w/ Concatenation	80.85	64.52
w/ Element-wise Product	81.80	58.46

Why use Emotion. We provide the detailed comparison of EmoForensics and SIMBA on the leave-one-out evaluation setup on FakeAVCeleb in Table 9. We highlight fluctuations in performance across individual Methods and Family splits relative to each model's average AUC in this evaluation setup.

Table 6. **Cross-manipulation performance comparison on FakeAVCeleb.** Performances are given as AUC. Best results are highlighted in **bold**, while the second-best are underlined.

	Faceswap Split	FSGAN Split	Wav2Lip Split	RTVC Split	Face Animation Family	Lip Synthesis Family	Average
AVAD [12]	73.63	75.01	94.27	73.34	74.32	94.75	80.89
AVFF[1] [34]	84.07	79.14	95.86	99.38	75.30	82.91	86.11
SIMBA [19]	80.87	**100.00**	**99.98**	89.97	**88.21**	**100.00**	93.17
EMO-BOOSTed SIMBA	**84.47**	**100.00**	**99.98**	**100.00**	88.11	99.23	**95.30**

Table 7. **Cross-manipulation method split performance on DeepSpeak v2.** Performances are given as AUC. Both SIMBA and Emo-Boosted SIMBA outperform the other deepfake detectors in terms of average AUC. Best results are highlighted in **bold**, and second-best are underlined.

	Liveportrait Real Audio	Facefusion Real Audio	Latentsync Real Audio	Memo Real Audio	Hellomeme Real Audio	Diff2lip Real Audio	Latentsync Elevenlabs	Latentsync Speechify	Latentsync Playht	Memo Playht	Memo Elevenlabs	Memo Speechify	Diff2lip Playht	Diff2lip Elevenlabs	Diff2lip Speechify
AVAD [12]	47.81	45.10	33.05	50.25	67.52	34.95	38.27	31.54	38.49	85.76	86.20	86.46	41.42	46.35	41.19
AVFF[1] [34]	80.80	**89.76**	99.81	93.38	79.88	99.71	99.85	**99.91**	98.33	88.94	93.82	97.76	96.07	**99.99**	99.89
SIMBA [19]	98.94	69.19	**99.85**	**99.95**	**94.36**	**100.00**	99.85	99.66	**99.61**	**99.77**	**99.87**	**99.78**	**99.93**	99.74	**100.00**
EMO-BOOSTed SIMBA	**99.03**	68.58	99.60	99.90	93.17	99.96	**99.92**	99.71	99.57	99.64	99.74	99.75	99.87	99.90	99.79

Table 8. **Cross-manipulation family split performance on DeepSpeak v2.** Family split performances are shown next to the average performance that is calculated across the method(Table 7) and family splits. Performances are given as AUC. Best results are highlighted in **bold**, and second-best are underlined.

	Avatar Family	Faceswap Family	LipSync Family	Average
AVAD [12]	55.19	45.10	34.00	50.48
AVFF[1] [34]	84.36	**85.28**	**99.97**	93.75
SIMBA [19]	**95.37**	62.87	96.59	**95.30**
EMO-BOOSTed SIMBA	94.580	65.48	96.45	95.26

Table 9. **Stability Ablation**. Comparison of performance variation across manipulation splits in the cross-manipulation evaluation setup for EmoForensics and SIMBA. Values in brackets indicate the difference from the average AUC across all manipulation and family splits for a given method. Red values denote a drop in performance relative to the average AUC, while green values indicate an increase.

	Faceswap Split	FSGAN Split	Wav2Lip Split	RTVC Split	Face Animation Family	Lip Synthesis Family	Average
EmoForensics	70.98(+1.68)	68.85(-0.45)	70.17(+0.87)	73.00(+3.70)	69.26(-0.04)	63.54(-5.76)	69.30
SIMBA	88.28(-6.23)	100.00(+5.50)	100.00(+5.50)	90.89(-3.61)	87.86(-6.65)	100.00(+5.50)	94.50