”Supplementary Material for Teacher-Student Structure for Domain Adaptation in Ensemble Audio-Visual Video Deepfake Detection”

A. Extra Figures

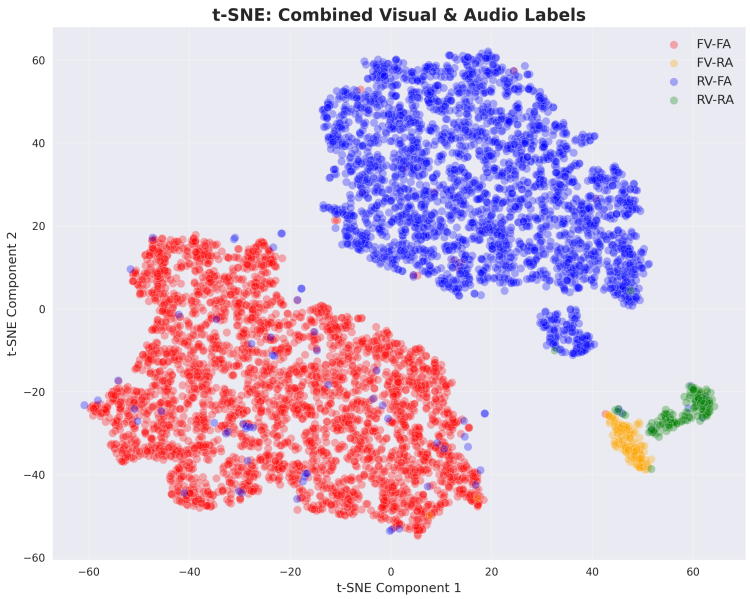

Fig. 1. t-SNE visualization of concatenated visual (V), audio (A), and audio-visual (AV) sub-network embeddings from the student model on the main dataset (FakeAVCeleb).

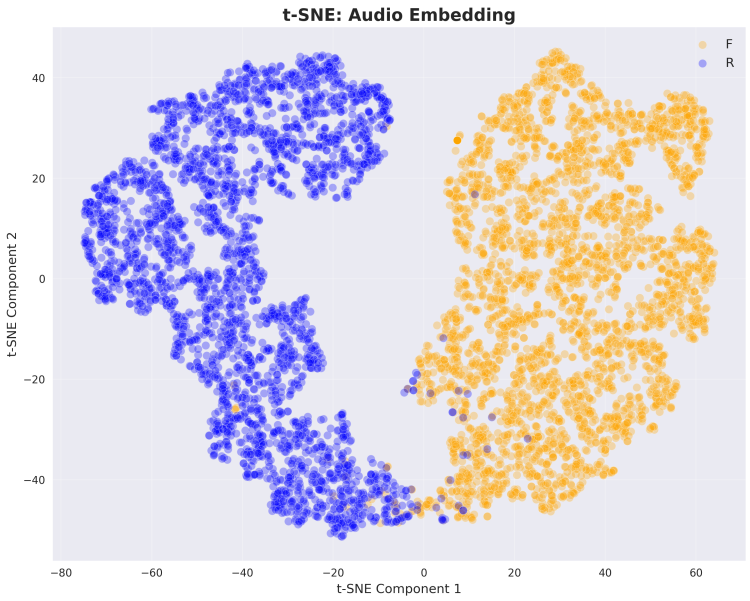

Fig. 2. t-SNE visualization of audio (A) sub-network embedding from the teacher model on the main dataset (FakeAVCeleb).

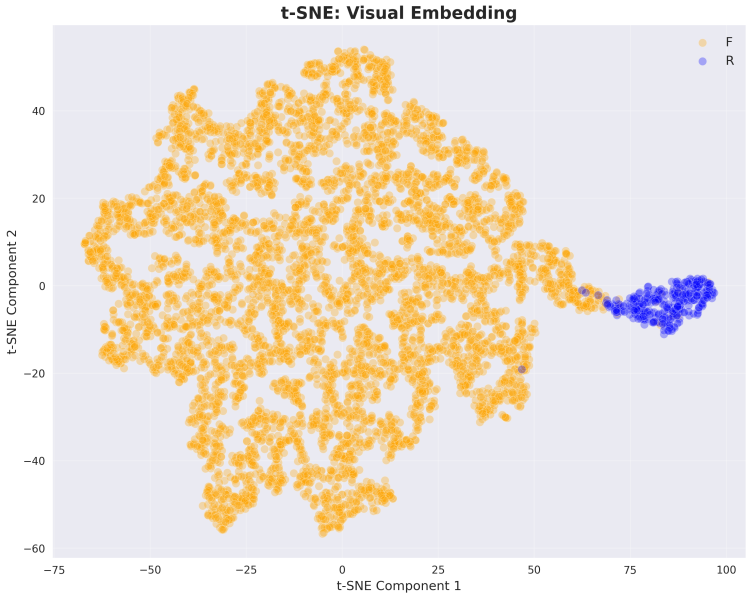

Fig. 3. t-SNE visualization of visual (V) sub-network embedding from the teacher model on the main dataset (FakeAVCeleb).

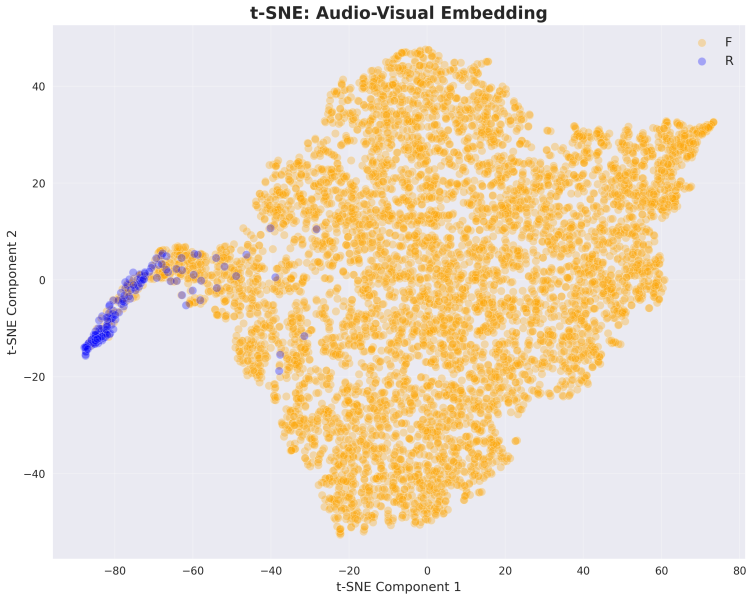

Fig. 4. t-SNE visualization of audio-visual (AV) sub-network embedding from the teacher model on the main dataset (FakeAVCeleb).

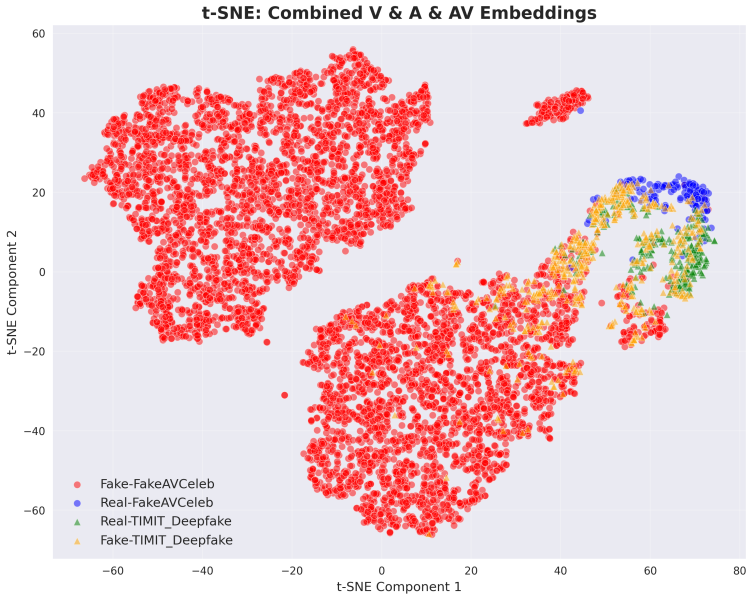

Fig. 5. t-SNE visualization of concatenated visual (V), audio (A), and audio-visual (AV) sub=network embeddings from the teacher model on the main dataset (FakeAVCeleb) and the cross-dataset (TIMIT_Deepfake). The visualization shows distinct clustering between real and fake samples across datasets.

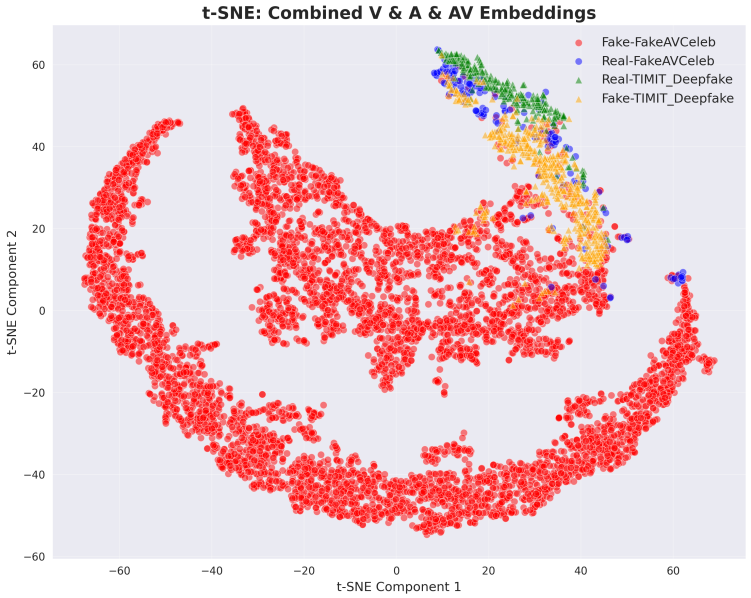

Fig. 6. t-SNE visualization of concatenated visual (V), audio (A), and audio-visual (AV) subnetwork embeddings from the student model on the main dataset (FakeAVCeleb) and the cross-dataset (TIMIT_Deepfake). The visualization shows distinct clustering between real and fake samples across datasets.

B. Extra Tables

TABLE I
MODEL SUB-NETWORKS AND THEIR CORRESPONDING DATASETS AND TASKS

Model Sub-Network	Dataset	Task
Audio-Visual [29]	LRS2 dataset	Binary synchronization classification
Visual	Faceforensics++	Binary deepfake classification
Audio	ASVspoof 2019	Binary deepfake classification

TABLE II
AVERAGE MODEL RESPONSE TIME ACROSS DIFFERENT DATASETS

	PolyGlotFake	DFDC	FakeAVCeleb	TIMIT_Deepfake
Avg time of model response for 7 clip inputs for each video (seconds)	0.18	0.2	0.17	0.135

TABLE III
PERFORMANCE COMPARISON OF TEACHER AND STUDENT MODEL SUB-NETWORKS ON POLYGLOTFAKE DATASET WITH DATASET III CONFIGURATION FOR STUDENT MODEL

Model	PolyGlotFake							
	Ensemble		Audio-Visual sub-network		Visual sub-network		Audio sub-network	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Teacher	93.25	97.28	93.99	97.15	66.22	93.80	77.79	91.49
Student	95.76	97.77	93.49	97.28	72.77	94.92	78.07	91.39

TABLE IV
PERFORMANCE COMPARISON OF TEACHER AND STUDENT MODEL SUB-NETWORKS ON FAKEAVCELEB DATASET WITH DATASET III CONFIGURATION FOR STUDENT MODEL

Model	FakeAVCeleb							
	Ensemble		Audio-Visual sub-network		Visual sub-network		Audio sub-network	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Teacher	99.33	99.88	99.03	99.71	99.09	99.64	99.65	99.99
Student	99.81	99.94	98.93	99.72	99.88	99.93	99.56	1

TABLE V

PERFORMANCE COMPARISON OF MODELS WITH DIFFERENT FUSION STRATEGIES ACROSS FOUR DATASETS: FAKEAVCELEB, DFDC, AND TIMIT_DEEPFAKE WHEN THE MODEL IS TRAINED ON FAKEAVCELEB.

Model	FakeAVCeleb		DFDC		TIMIT_Deepfake		PloyGlotFake	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Teacher (Split)	99.33	99.88	**76.11**	**67.33**	72.25	81.71	**93.25**	**97.27**
Teacher (Joint)	**99.81**	**99.94**	39.93	63.19	65.91	65.91	69.75	89.61
Teacher (Attention)	98.66	99.81	60.47	42.55	**77.30**	**88.42**	66.39	86.86