Unsupervised Multimodal Deepfake Detection Through Explicit Intra-Modal and Cross Modal Inconsistency Discovery

USC Viterbi Information Sciences Institute

Mulin Tian, Mahyar Khayatkhoei, Joe Mathai, Wael AdbAlmageed

BMVC 2025

1.Motivation

Deepfake videos present an increasing threat to society with potentially negative impact on criminal justice, democracy, and personal safety and privacy. Meanwhile, detecting deepfakes, at scale, remains a very challenging task that often requires labeled training data from existing deepfake generation methods. Further, even the most accurate supervised deepfake detection methods do not generalize to deepfakes generated using new generation methods. In this paper, we propose a novel unsupervised method for detecting deepfake videos by explicitly discovering intra-modal and cross-modal inconsistencies between video segments. The fundamental hypothesis behind our proposed method is that motion or identity inconsistencies are inevitable in motion-transfer deepfake videos. We provide the first mathematical and empirical evidence to support this hypothesis, hence justifying the generalizability of our method. Our proposed method outperforms prior state-of-the-art unsupervised deepfake detection methods on the challenging FakeAVCeleb dataset, and also has several additional advantages: it is scalable because it does not require pristine (real) samples for each identity during inference and therefore can apply to arbitrarily many identities; generalizable because it is trained only on real videos and therefore does not rely on any particular deepfake method; reliable because it does not rely on any likelihood estimation in high dimensions; and explainable because it can pinpoint the exact location of modality inconsistencies.

2.Intra-Modal and Cross-Modal Deepfake Detection

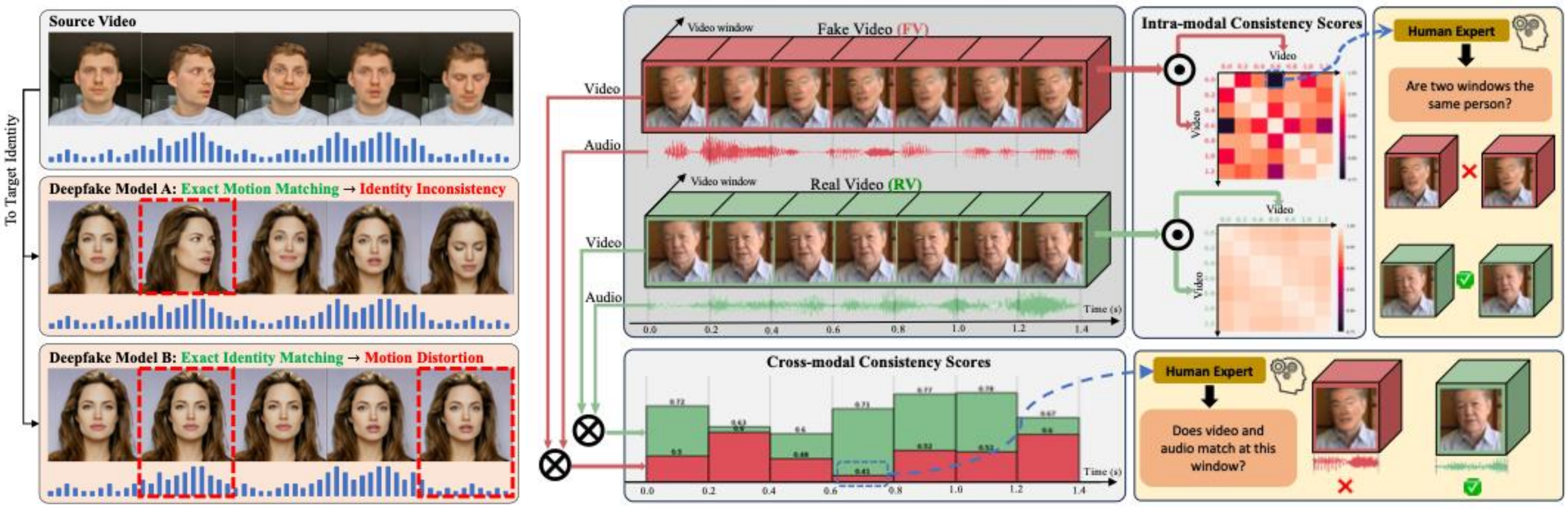
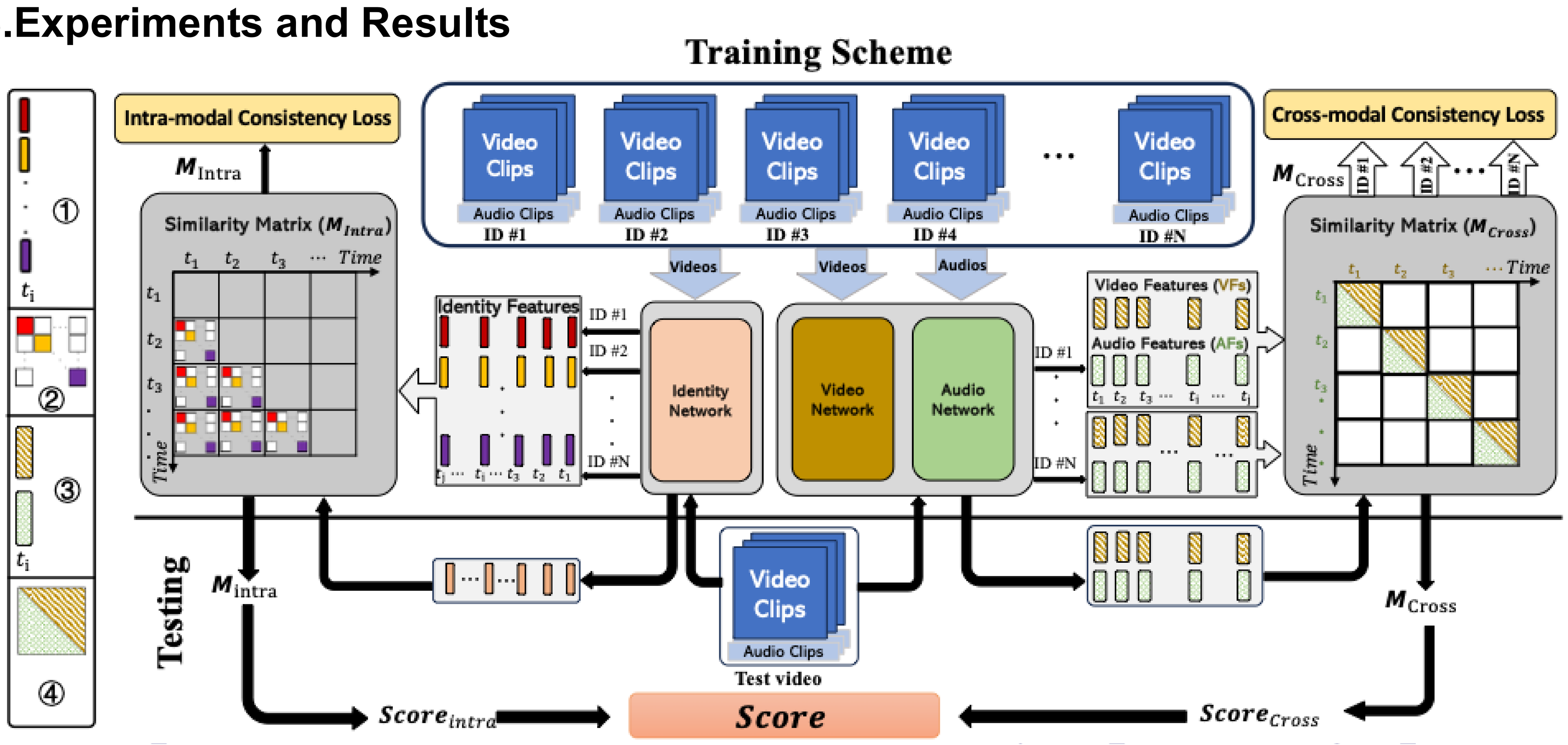


Figure 1. Left: Illustration of the inevitable trade-off in deepfake generation—matching the source motion precisely leads to **identity inconsistency**, while preserving the target identity results in **motion distortion** and audio-visual desynchronization. **Right**: Our proposed **unsupervised multimodal detection framework** explicitly identifies these **intra-modal (identity)** and **cross-modal (audio-visual)** inconsistencies within a single video, enabling scalable and explainable deepfake detection.

3.Experiments and Results



	Methods	FakeAVCeleb HQ	FakeAVCeleb LQ	KoDF Attacked
Supervised	Seferbekov[46]	**98.6**	**61.7**	**61.5**
	FTCN[51]	84.0	37.6	58.3
	LipForensics [21]	97.6	58.3	54.8
	Real Forensics[22]	88.3	52.9	55.5
	MDS-based FD[5]	64.7	61.1	55.4
	Joint AV[52]	55.1	55.2	47.1
Unsupervised	ICT [14]	68.2	66.9	61.5
	ICT-Ref[14]	71.9	71.2	78.0
	ID-Reveal [10]	70.2	70.8	73.4
	POI-Forensics [8]	94.1	**94.4**	80.5
	Ours (Intra-Cross-modal)	**95.5**	85.9	**84.2**±1.1

	Method	Modality	Pretrained Dataset	Catogory RVFA AP	RVFA AUC	FVRA-WL AP	FVRA-WL AUC	FVFA-FS AP	FVFA-FS AUC	FVFA-GAN AP	FVFA-GAN AUC	FVFA-WL AP	FVFA-WL AUC	AVG-FV AP	AVG-FV AUC
Supervised	Xception [45]	V	ImageNet	-	-	88.2	88.3	92.3	93.5	67.6	68.5	91.0	91.0	84.8	85.3
	LipForensics [21]	V	LRW	-	-	**97.8**	97.7	99.9	99.9	61.5	68.1	98.6	98.7	89.4	91.1
	AD DFD [52]	AV	Kinetics	74.9	73.3	97.0	97.4	99.6	99.7	58.4	55.4	**100.**	**100.**	88.8	88.1
	FTCN [51]	V	-	-	-	96.2	97.4	**100.**	**100.**	77.4	78.3	95.6	96.5	92.3	93.1
	RealForensics [22]	V	LRW	-	-	88.8	93.0	99.3	99.1	99.8	99.8	93.4	96.7	95.3	97.1
	AVFF [40]	AV	LRS3 & FakeAvCeleb	**93.3**	**92.4**	94.8	**98.2**	**100.**	**100.**	**99.9**	**100.**	99.4	99.8	**98.5**	**99.5**
Unsupervised	AVBYOL [19]	AV	LRW	50.0	50.0	73.4	61.3	88.7	80.8	60.2	33.8	73.2	61.0	73.9	59.2
	VQ-GAN [16]	V	LRS2	-	-	50.3	49.3	57.5	53.0	49.6	48.0	62.4	56.9	55.0	51.8
	A-V Anomaly [17]	AV	LRS2	62.4	71.6	93.6	93.7	95.3	95.8	94.1	94.3	93.8	94.1	94.2	94.5
	A-V Anomaly [17]	AV	LRS3	70.7	80.5	91.1	93.0	91.0	92.3	91.6	92.7	91.4	93.1	91.3	92.8
	Ours (Intra-modal)	V	VoxCeleb2	-	-	65.14	67.99	86.11	86.65	90.08	90.65	63.28	66.15	76.15	77.86
	Ours (Cross-modal)	AV	VoxCeleb2	**99.68**	**99.65**	94.56	95.89	93.52	95.66	92.16	94.58	**95.29**	**96.25**	93.88	95.62
	Ours (Intra-Cross-modal)	AV+V	VoxCeleb2	98.49	99.41	**94.92**	**95.96**	**96.90**	**97.71**	**96.67**	**97.59**	94.84	95.99	**95.74**	**96.81**

Figure 2 & Tables 1–2. ***Left-top***: Overview of the proposed **intra-modal** and **cross-modal consistency** framework, which jointly learns to detect **identity** and **motion inconsistencies** within and across video–audio segments. ***Right-top***: Generalization studies showing that our method remains robust under **video/audio compression** and **adversarially attacked KoDF** datasets, outperforming prior supervised and unsupervised methods. ***Bottom***: Quantitative comparison on **FakeAVCeleb**, where our combined **Intra-Cross-modal** method achieves new state-of-the-art performance among unsupervised detectors and approaches supervised results. **All results are evaluated using Area Under the ROC Curve (AUC)** and **Average Precision (AP)** metrics.