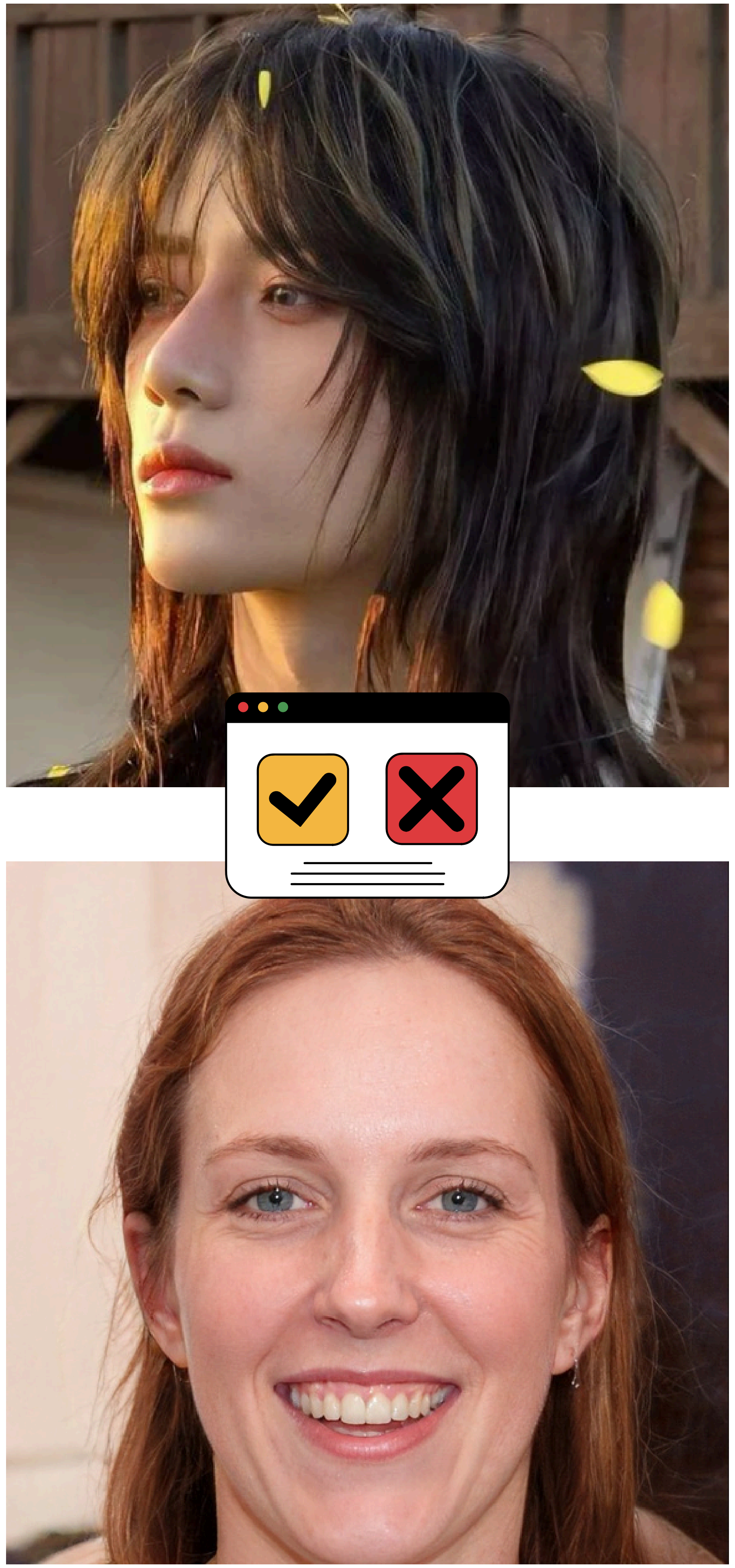
Deepfakes & ILLUSION

A PRIMER AND RESEARCH UNDERTAKEN AT IITJ

Akanksha Singh
23 January 2024

CeRAI
Centre for Responsible AI

Motivation:
To curb the malicious use of Generative AI and deepfakes.



AGENDA

- Overview
- Deepfake Generation
- Deepfake Detection
- ILLUSION

OVERVIEW

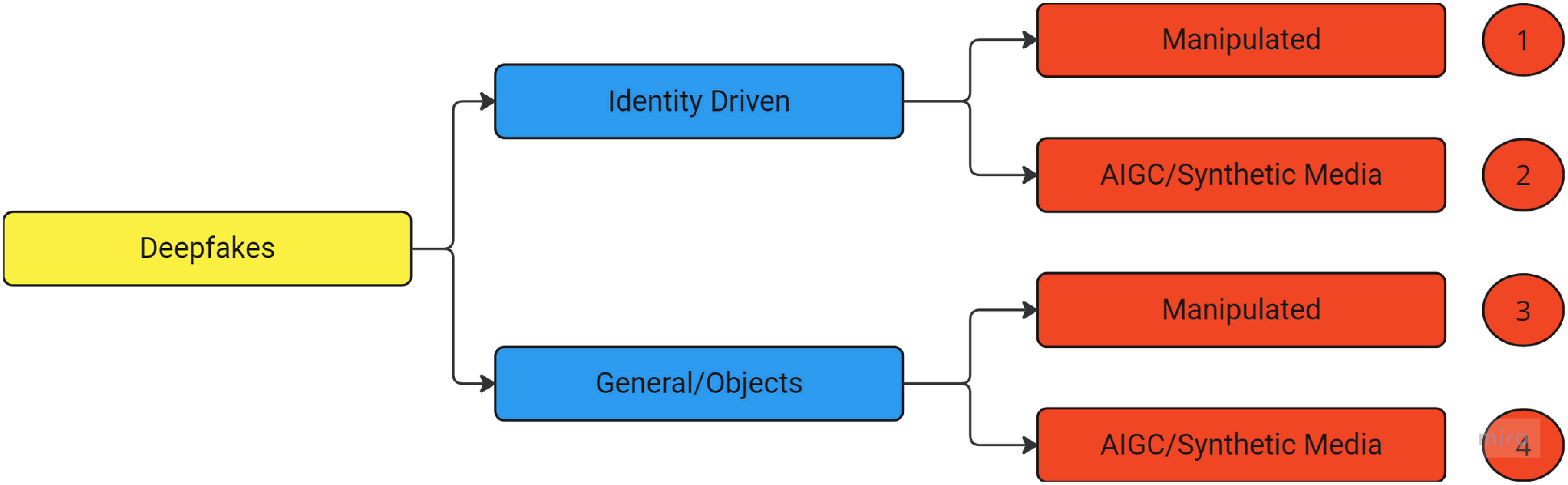


Figure: Deepfakes based on whether content is 1. based on **human biometrics** and 2. **partial or fully** synthetic.

DEEPFAKE GENERATION

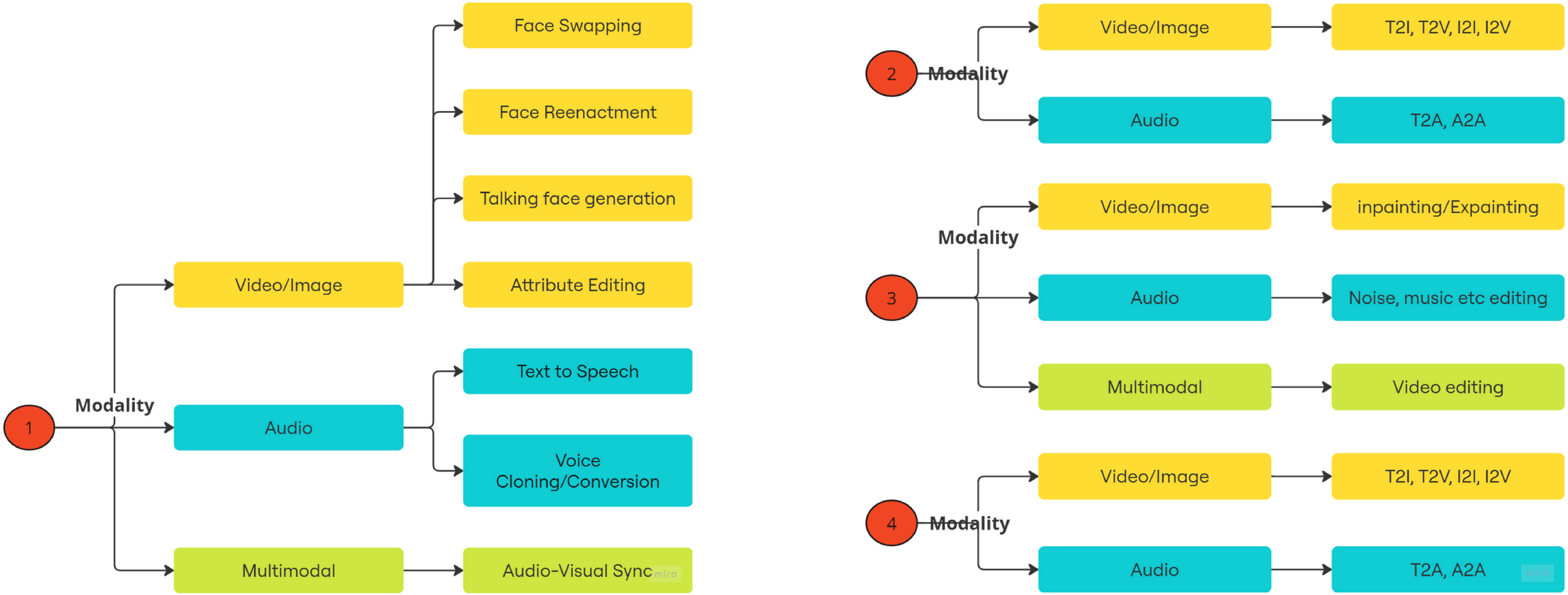


Figure: 1. ID-Driven Partial Manipulations, 2. ID-Driven Synthetic Media, 3. General Partial Manipulations, 4. General Synthetic Media

Source: A Comprehensive Survey with Critical Analysis for Deepfake Speech Detection, https://github.com/Daisy-Zhang/Awesome-Deepfakes-Detection, Deepfake generation and detection: A benchmark and survey

DEEPFAKE DETECTION

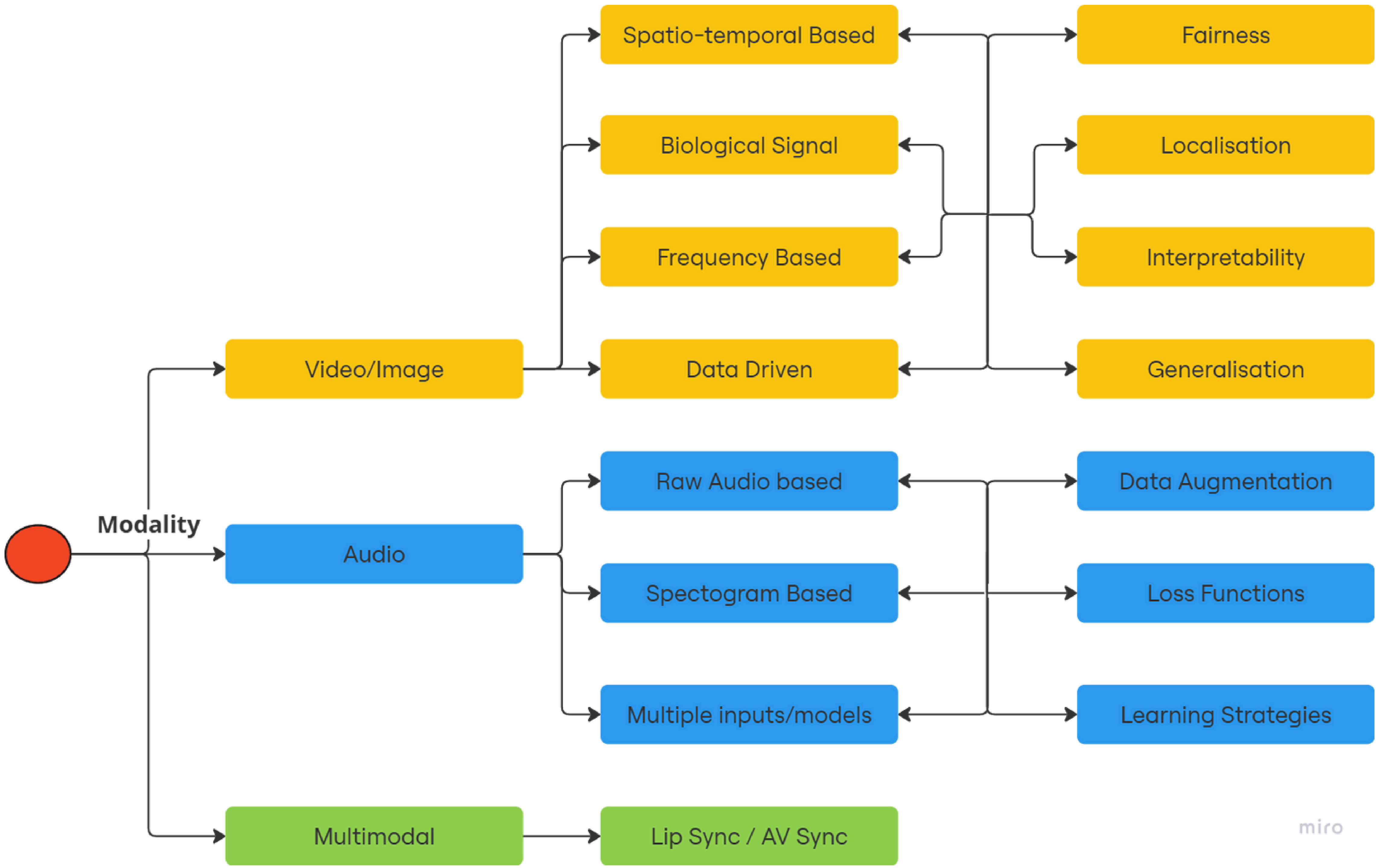


Figure: Common techniques used for each modality and additional research directions.

Source: A Comprehensive Survey with Critical Analysis for Deepfake Speech Detection, https://github.com/Daisy-Zhang/Awesome-Deepfakes-Detection, Deepfake generation and detection: A benchmark and survey

INTRODUCTION

Problem Statement

The purpose of the dataset is to aid in the creation of multimodal deepfake detection algorithms that are robust to all forms of fake media and unified across all three modalities, are bias-free and imperceptible to human eyes.

1

Research Gaps

Unimodal

Most SOTAs are unimodal

Variation

Exhaustive list of models, video length, quality of sync

Bias

Sex and skin-type biases

RELATED WORKS

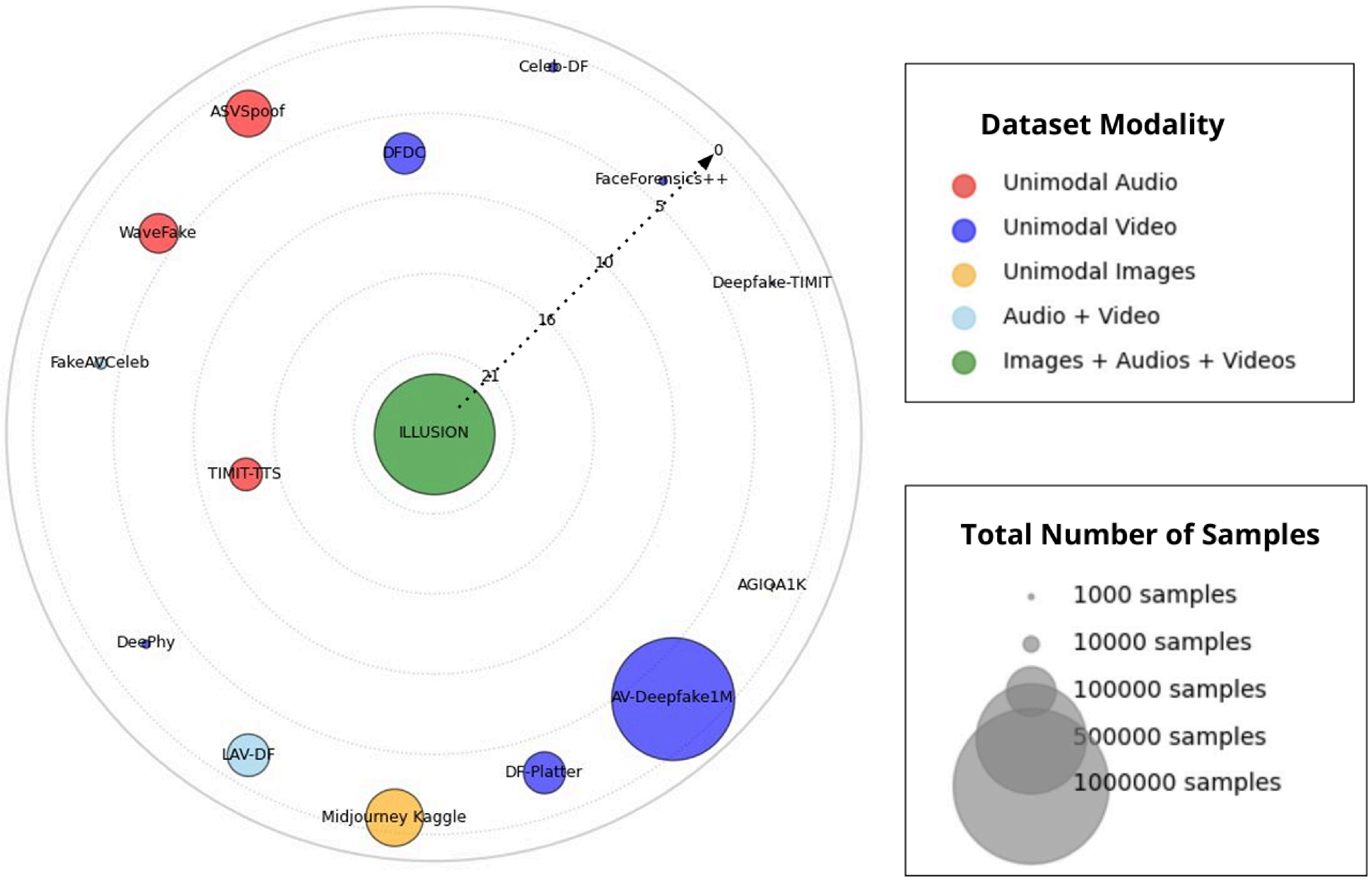


Figure: Comparative analysis of the proposed dataset with existing ones based on modalities, size, and manipulations.

Source: ILLUSION: Unveiling Truth with a Comprehensive Multi-Modal, Multi-Lingual Deepfake Dataset Accepted at ICLR 2025

ILLUSION DATASET

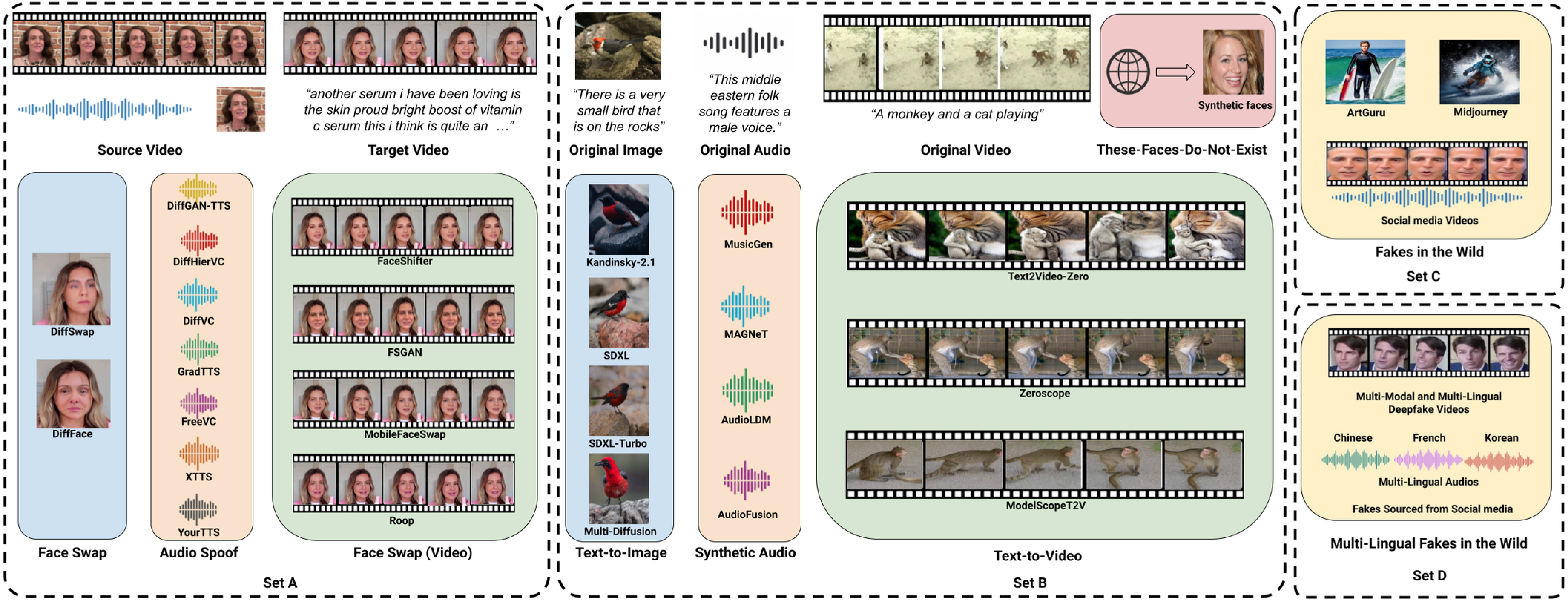


Figure: Pictoral representation of sets and techniques used in each of the proposed dataset.

Source: ILLUSION: Unveiling Truth with a Comprehensive Multi-Modal, Multi-Lingual Deepfake Dataset Accepted at ICLR 2025

QUANTATIVE AND QUALITATIVE ANALYSIS

28
techniques

4
sets

139740
real samples

27244
fake audio

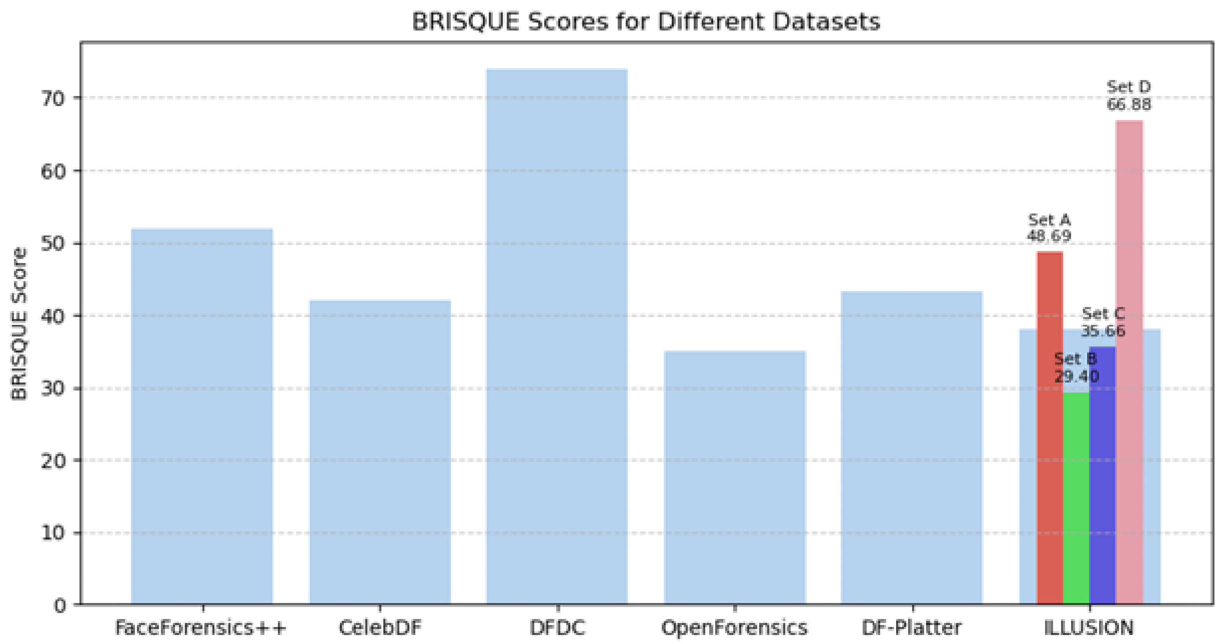
299454
fake videos

905548
fake images

1371986
total samples

Figure: Collated samples of techniques used.

Table: Visual quality comparison of existing datasets with our proposed dataset.



Source: ILLUSION: Unveiling Truth with a Comprehensive Multi-Modal, Multi-Lingual Deepfake Dataset Accepted at ICLR 2025

GENERATION PIPELINE: SET A

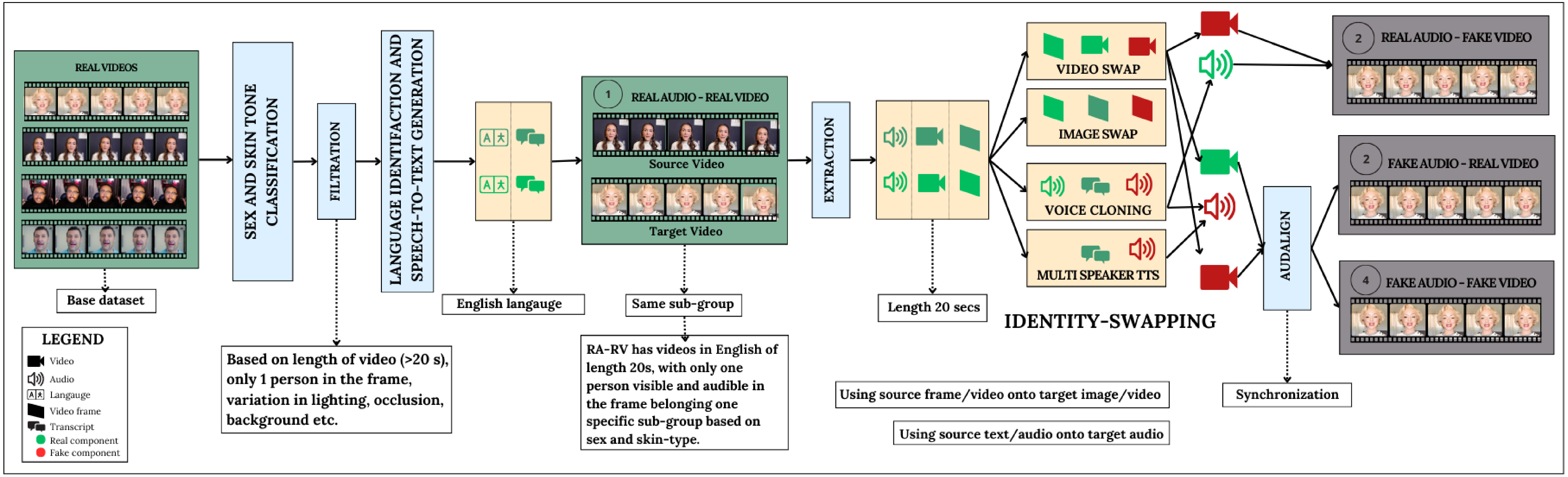


Figure: Creation of each class label of Set A: ID-Driven Partial Manipulation

Source: ILLUSION: Unveiling Truth with a Comprehensive Multi-Modal, Multi-Lingual Deepfake Dataset Accepted at ICLR 2025

EXPERIMENT 1

RQ
RQ1

Can we detect identity-aware swaps?

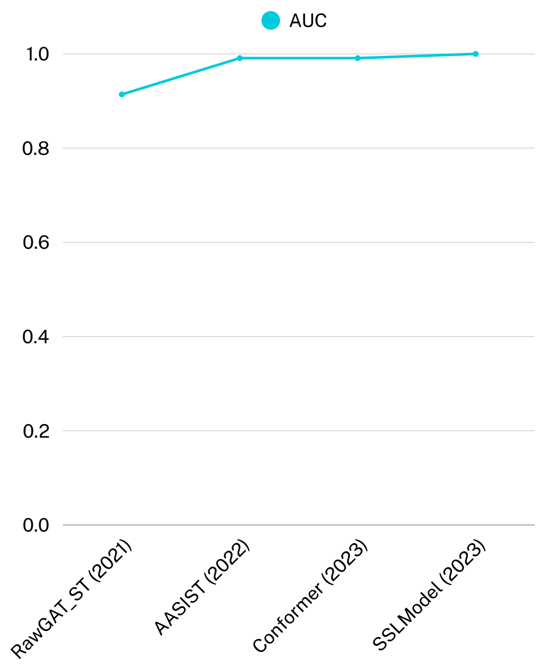
Protocol
Detection

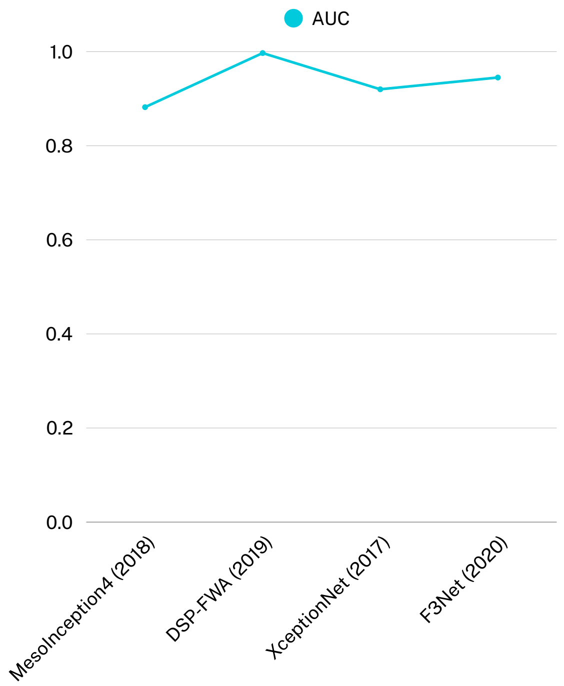
- Visual: Additional 18 videos per sub-group; 500 frames per video in train set and 240 frames in test for balancing.
- Audio: Balancing through weights in loss function.

Hypothesis

The model should perform better when trained and tested on our dataset because of the variation in techniques, quantity and balancing.

Results





Analysis

Models trained on our dataset show good learning of both real and fake classes.

Table: Results of Unimodal baseline experiments done on identity-aware swaps.

EXPERIMENT 2

RQ
RQ2

Are the state-of-the-art detection algorithms sufficiently robust for deployment in real-world scenarios?

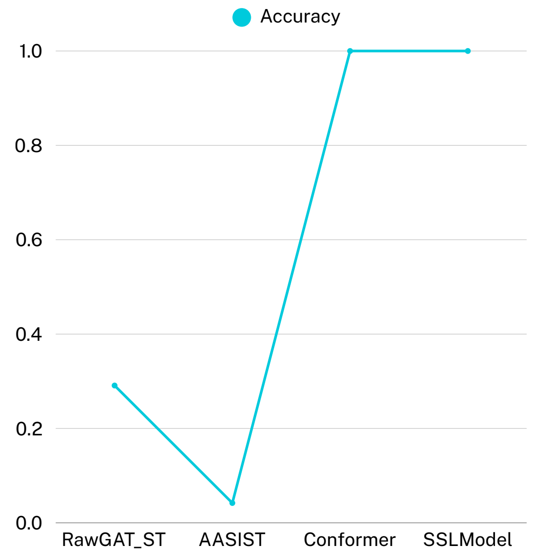
Protocol
Real World Data

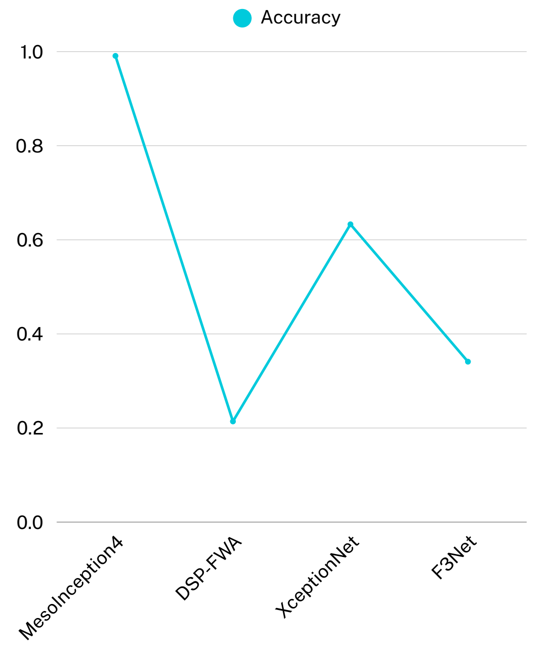
Use real world samples collected as a part of the project.

Hypothesis

Models trained on our dataset are robust against real-world samples with unknown manipulation techniques because of the variation introduced for deepfake detection.

Results





Analysis

Not all models are readily deployable for detection in real-world scenario. Pertinent to develop more generalisable detection algorithm.

Table: Results of models trained on identity-aware swaps and tested on real world samples.

EXPERIMENT 3

RQ
RQ3

Is zero-shot/zero-day detection possible?

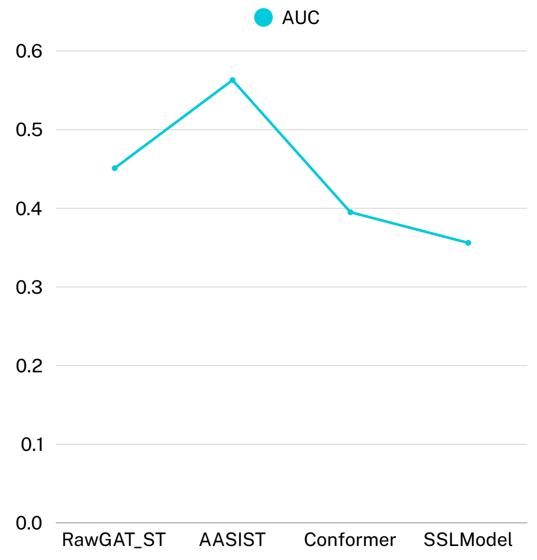
Protocol
Zero-day Attack

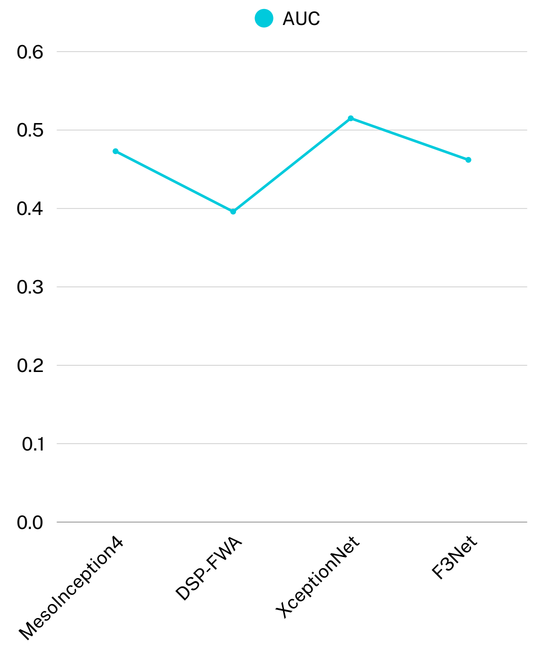
Use real world samples collected as a part of the project.

Hypothesis

Models trained on our dataset should not perform very well (>0.9 AUC) given the stark generation difference between the train and test sets.

Results





Analysis

The models trained on identity-aware swaps is not generalisable and capable of detection of entirely synthetic media.

Table: Results of models trained on Set A and tested on Set B of ILLUSION dataset

EXPERIMENT 4

RQ
RQ4

Are the state-of-the-art detection algorithms sufficiently robust against quality variation introduced during transmission in real world scenario?

Protocol
Augmentations

Perform c23 and c40 compression for visual part of the dataset.

Hypothesis

Compression mimics real-world data better. Models trained on our dataset perform well under compression tests.

Results

	Tested On	AUC		
Trained On	Model	Raw	c23	c40
	MesoInception4	0.882	0.849	0.73
	DSP-FWA	0.997	0.997	0.874
	XceptionNet	0.92	0.906	0.843
Raw	F3Net	0.945	0.927	0.852
	MesoInception4	0.917	0.911	0.825
	DSP-FWA	0.995	0.996	0.93
	XceptionNet	0.929	0.917	0.859
c23	F3Net	0.949	0.938	0.856
	MesoInception4	0.738	0.814	0.869
	DSP-FWA	0.869	0.838	0.978
	XceptionNet	0.847	0.851	0.879
c40	F3Net	0.84	0.862	0.884

Analysis

The models trained on identity-aware swaps is not generalisable and capable of detection of entirely synthetic media.

Table: Results of models trained on Set A and tested on Set B of ILLUSION dataset

EXPERIMENT 5

RQ
RQ5

Can we identify the source model of the deepfake?

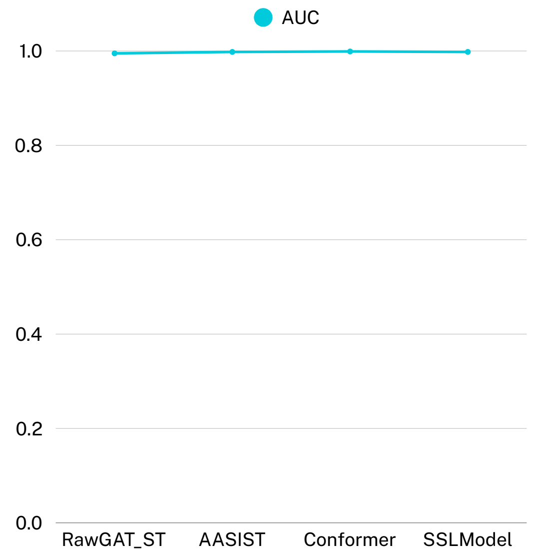
Protocol
Attribution

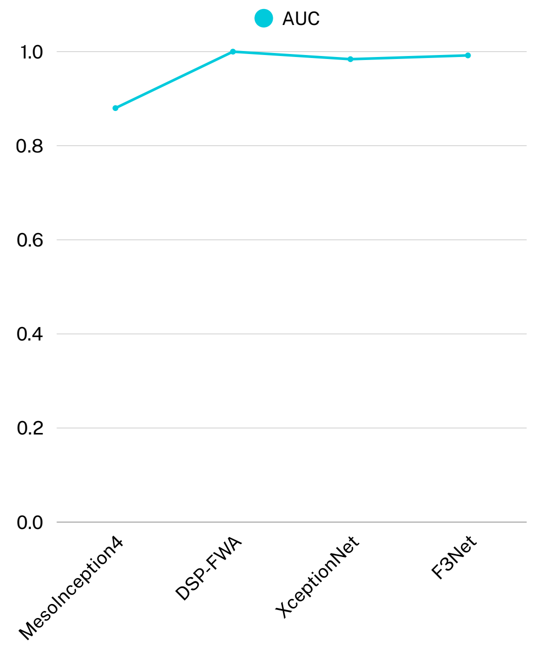
Show equal number of samples for each generation technique.

Hypothesis

The models in our dataset learn artifacts unique to each generation technique well and variation in dataset enables more generalisable learning for detection.

Results





Analysis

Models learn to distinguish artifacts of each generation technique.

Table: Results of generation model attribution of identity-aware swaps

Thank you!

DO YOU HAVE ANY QUESTIONS?

Akanksha Singh
23 January 2024