ICASSP 2024 KOREA

OPEN-SET DEEPFAKE DETECTION TO FIGHT THE UNKNOWN

Michael Diniz[1,3] **Anderson Rocha[1,2]**

[1]Recod.ai Lab – Reasoning for Complex Data

[2]Institute of Computing – Campinas State University

[3]Federal Institute of São Paulo

UNICAMP

INSTITUTE OF COMPUTING

recod.ai

FEDERAL INSTITUTE OF SÃO PAUO

UNICAMP

INSTITUTE OF COMPUTING

DIFFICULTIES IN DETECTING DEEPFAKE

recod.ai

FEDERAL INSTITUTE OF SÃO PAUO

NO SEMANTICS

HIGH QUALITY

HIGH Diversity

DeepBrain

Reface

Deepfakes Web

AttGAN

StyleGAN

StarGANv2

STGAN

FaceApp

RSGAN

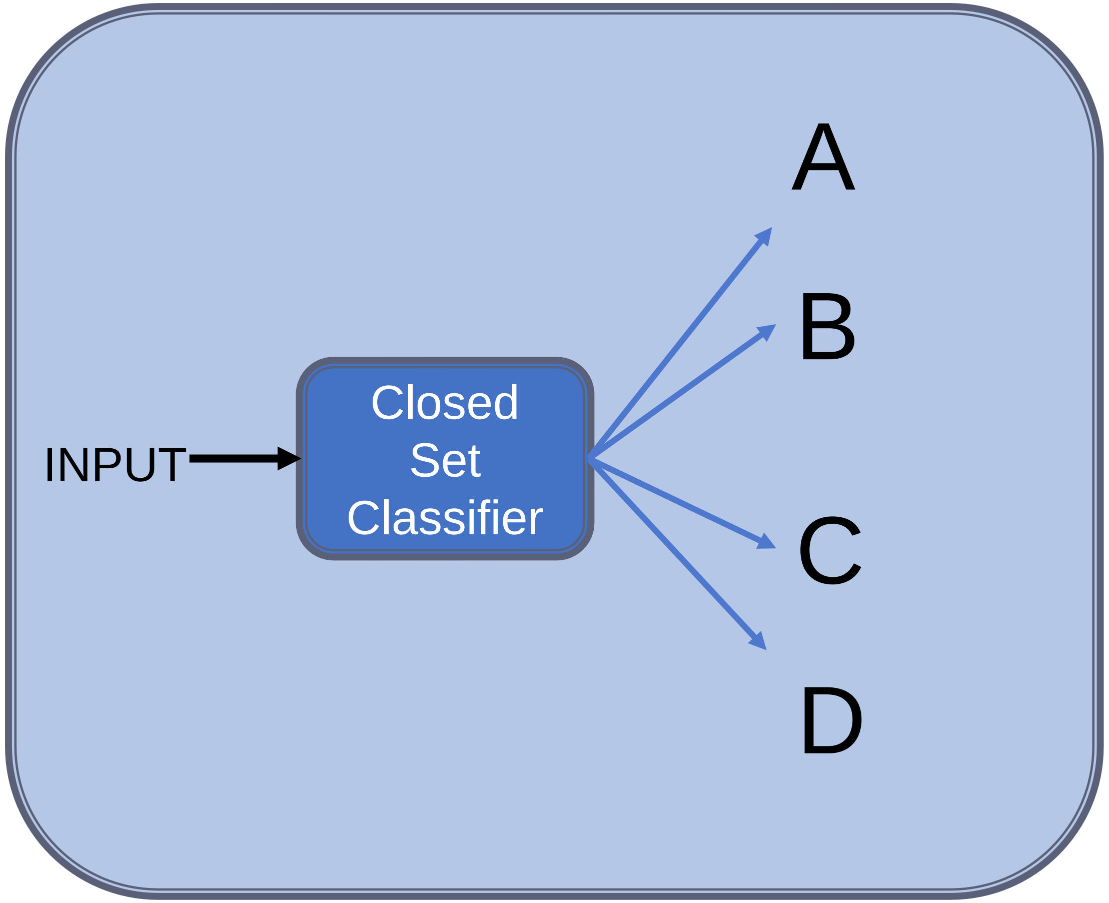
LipGAN

> 42 user-friendly deepfake tools

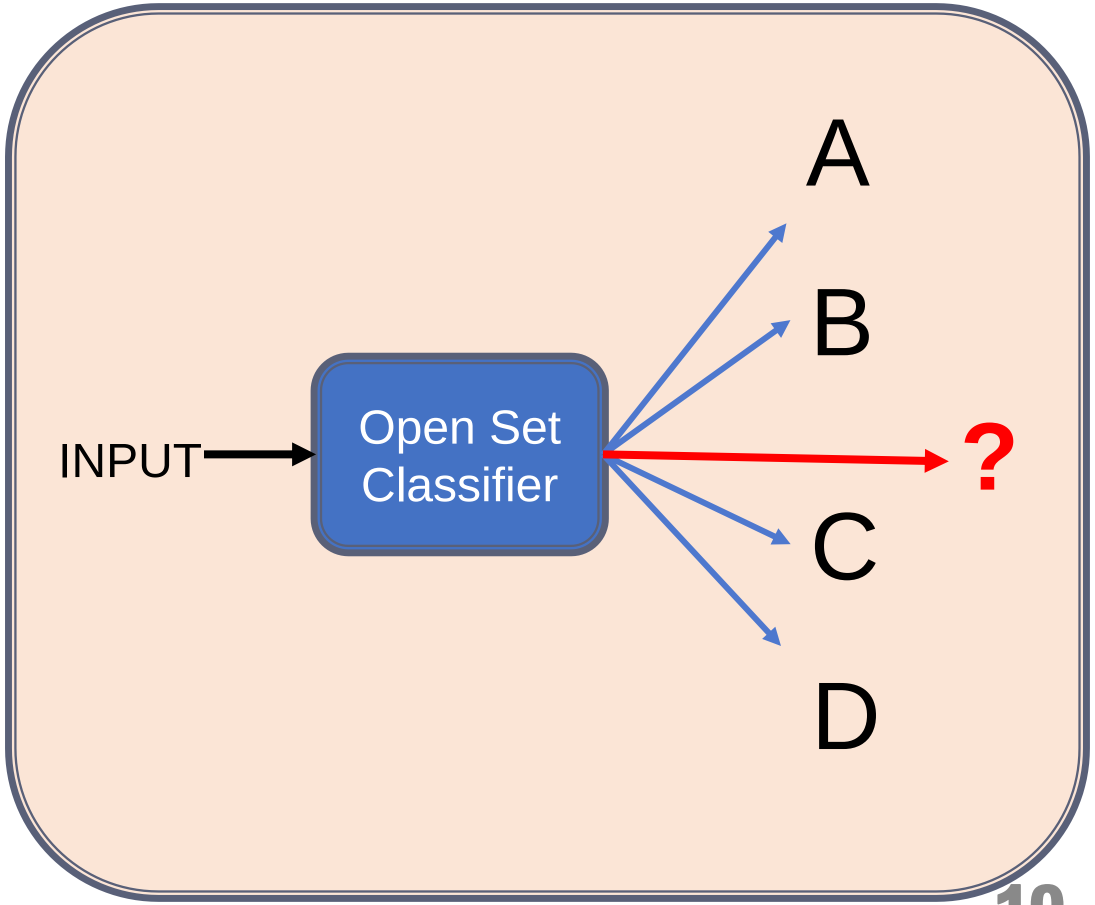
https://www.homesecurityheroes.com/state-of-deepfakes/

CLOSED SET X OPEN SET

Closed Set Classification

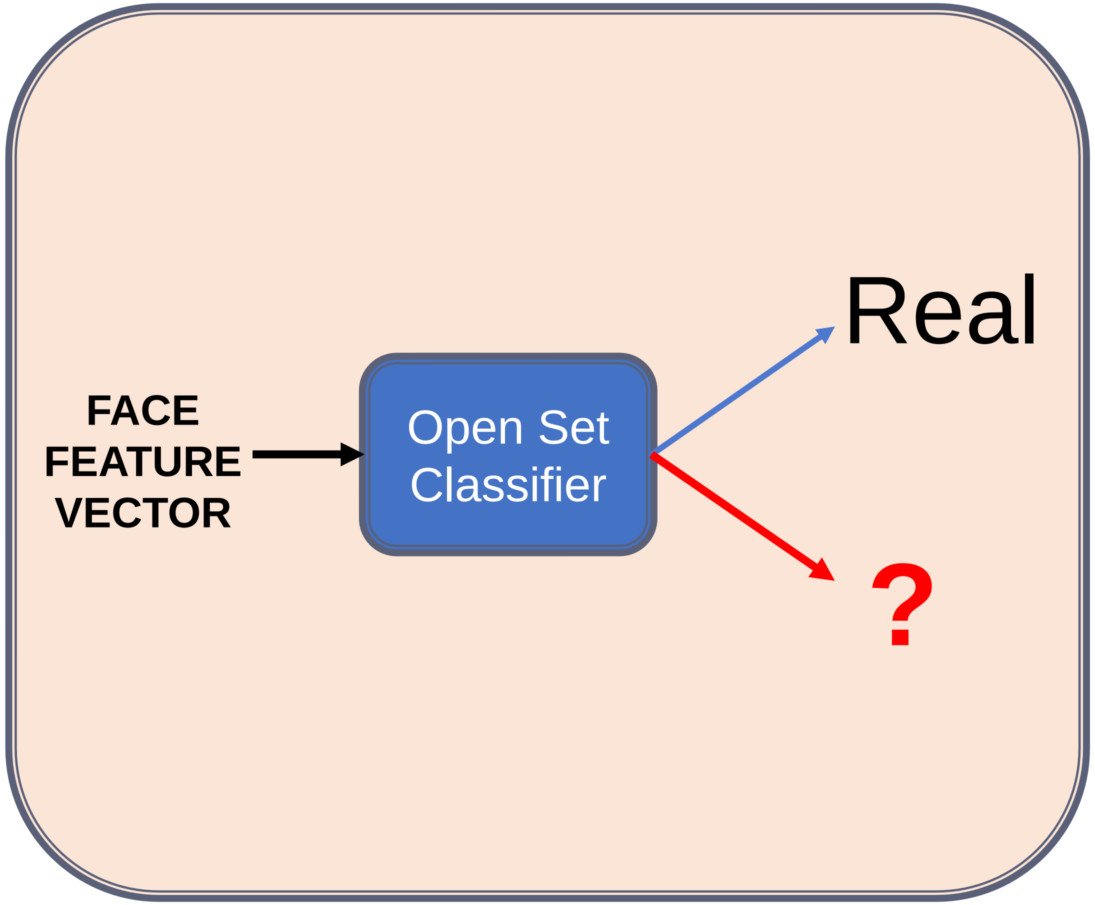


Open Set Classification



CLOSED SET **X** OPEN SET

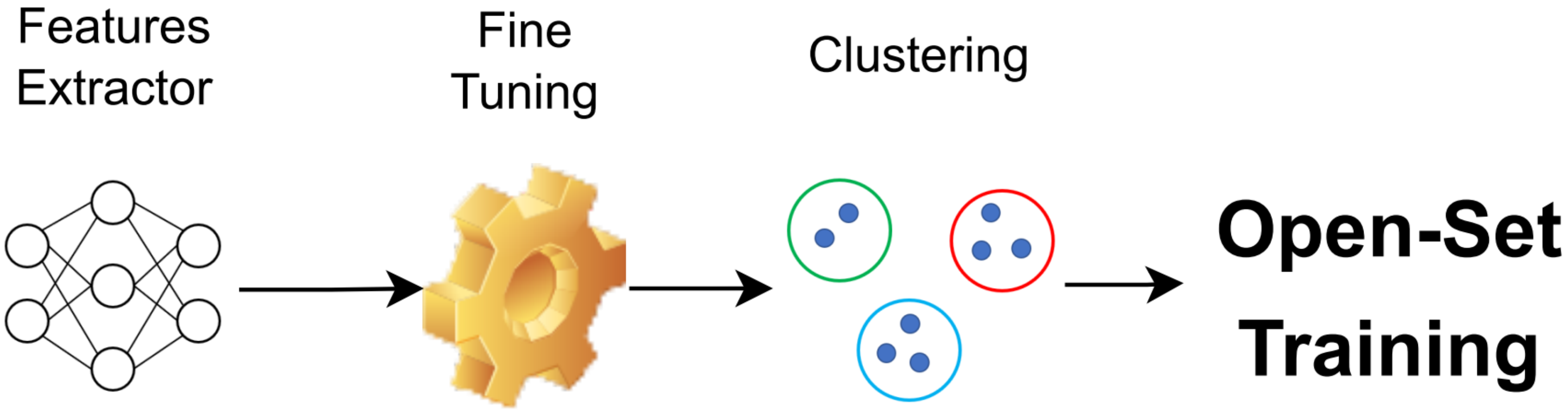
Open Set Classification



OBJECTIVE

To propose and investigate an open-set approach for deepfake detection in images.

METHODOLOGY



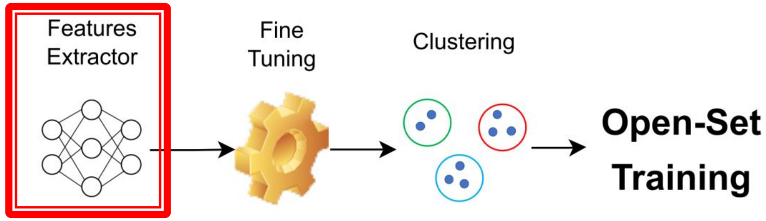
Features Extractor

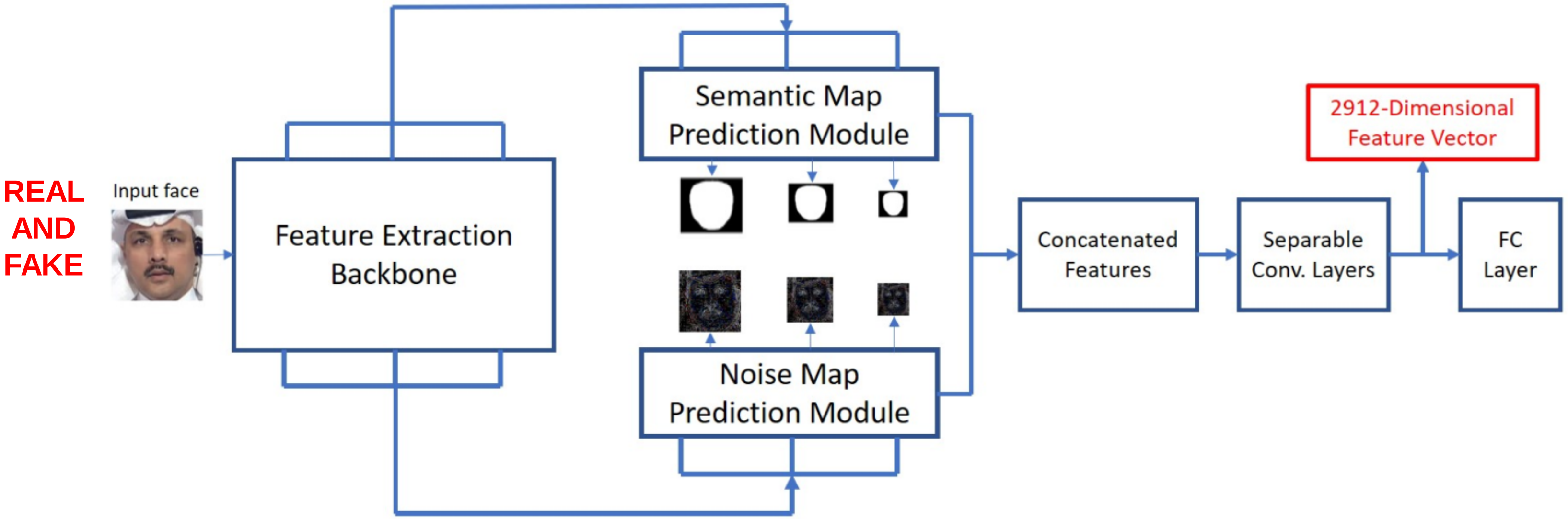
Fine Tuning

Clustering

Open-Set Training

METHODOLOGY: FEATURE EXTRACTOR




REAL AND FAKE
Input face
Feature Extraction Backbone
Semantic Map Prediction Module
Noise Map Prediction Module
Concatenated Features
Separable Conv. Layers
2912-Dimensional Feature Vector
FC Layer

Baseline

$$L = L_c + \lambda_1 L_n + \lambda_2 L_b$$

HTL-C

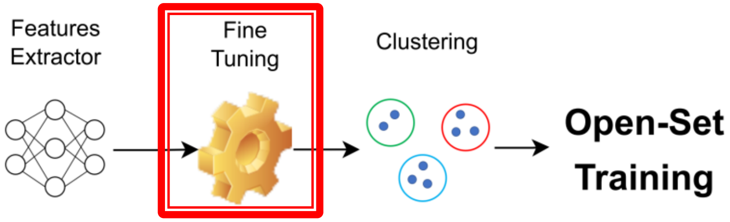
$$L = L_c + \lambda_1 L_n + \lambda_2 L_b + \lambda_3 L_T$$

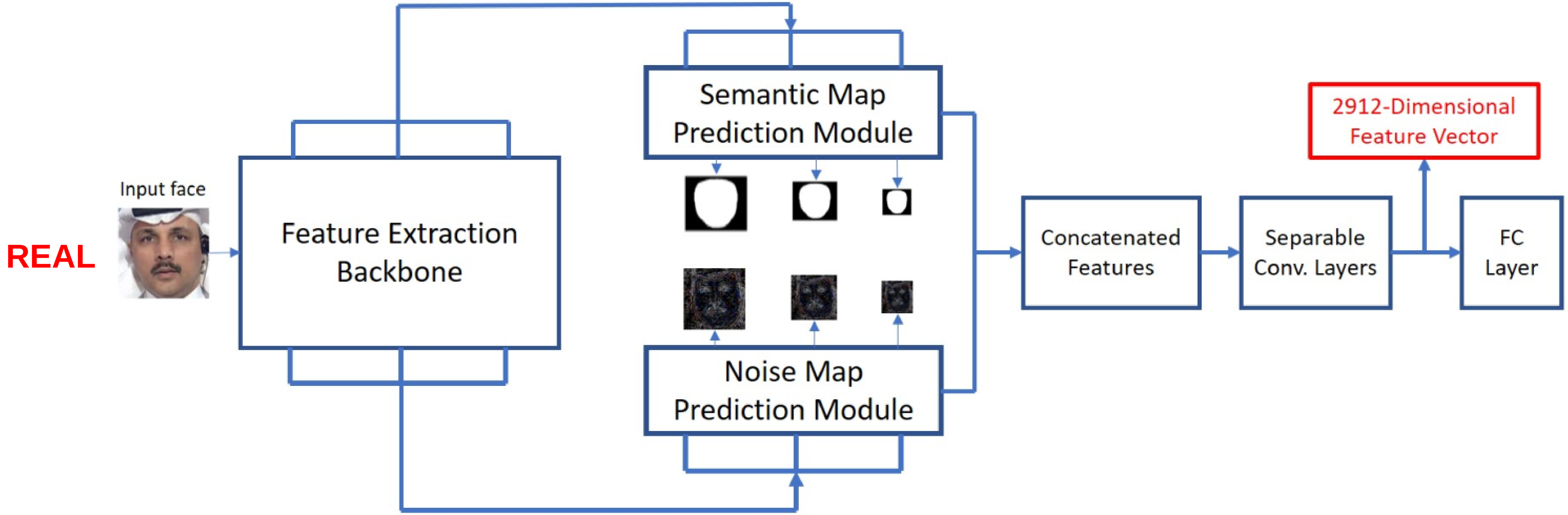
HTL-NC

$$L = \lambda_1 L_n + \lambda_2 L_b + \lambda_3 L_T$$

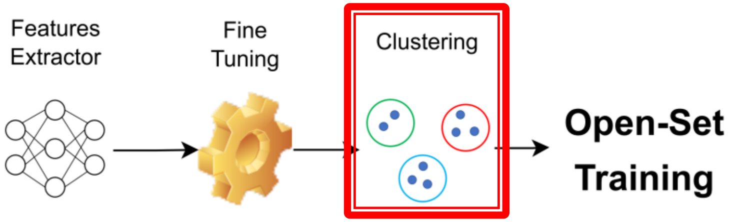
Chenqi Kong, Baoliang Chen, Haoliang Li, Shiqi Wang, Anderson Rocha, and Sam Kwong, “**Detect and locate: Exposing face manipulation by semantic-and noise-level telltales**,” *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1741–1756, 2022

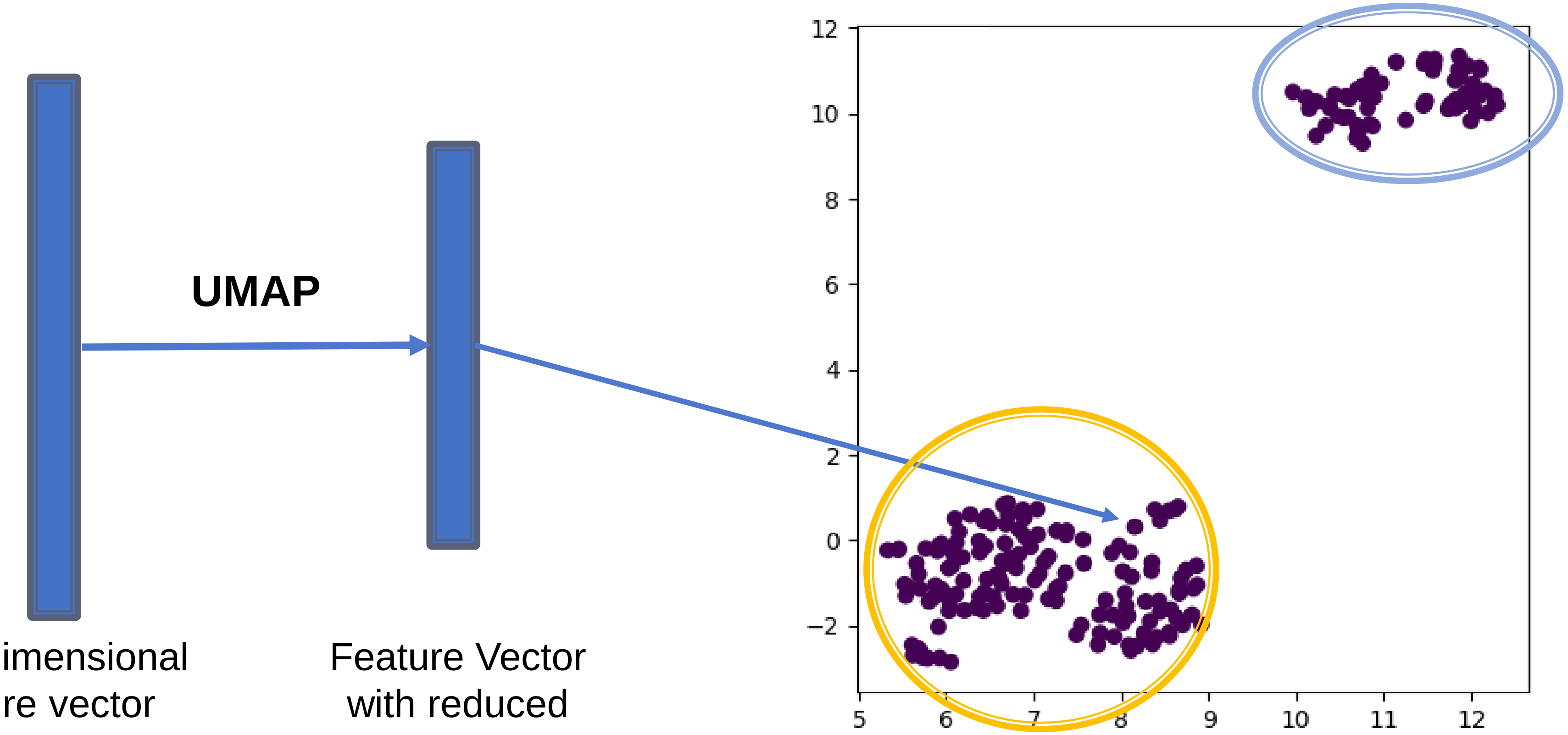
METHODOLOGY: FINE TUNING




REAL
Input face
Feature Extraction Backbone
Semantic Map Prediction Module
Noise Map Prediction Module
Concatenated Features
Separable Conv. Layers
2912-Dimensional Feature Vector
FC Layer

METHODOLOGY: CLUSTERING

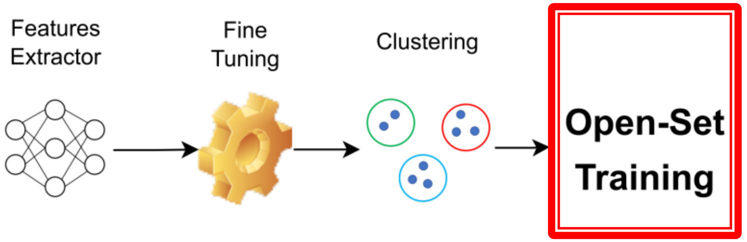


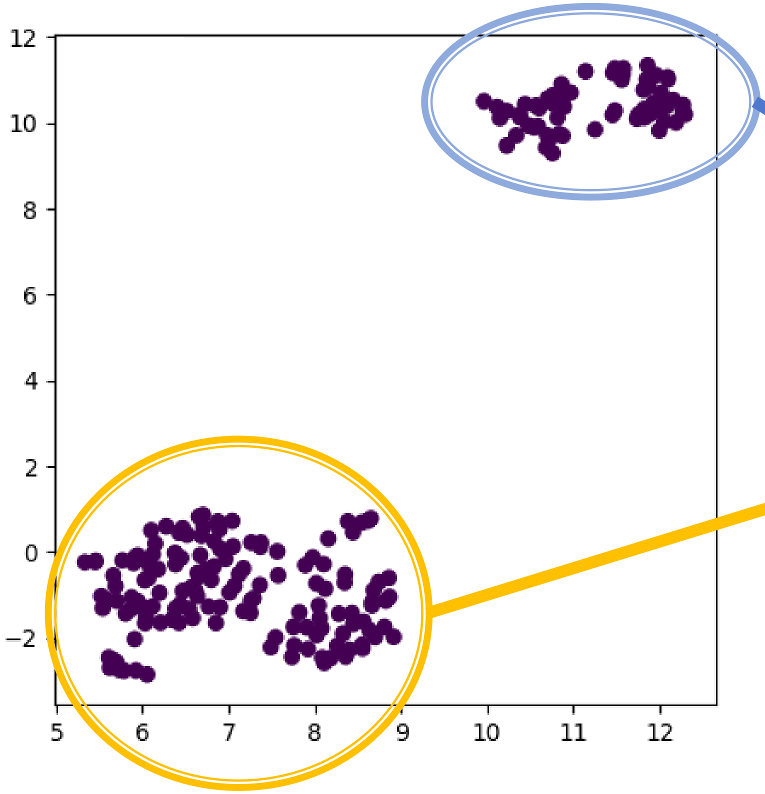


2912 dimensional feature vector

Feature Vector with reduced dimension

METHODOLOGY: OPEN-SET TRAINING





- One Class SVM
- Extreme Valued Machine
- Isolation Forest

EXPERIMENTS AND RESULTS

DATA SET

Feature Extractor Training

Full FaceForensics ++ C23 and C40

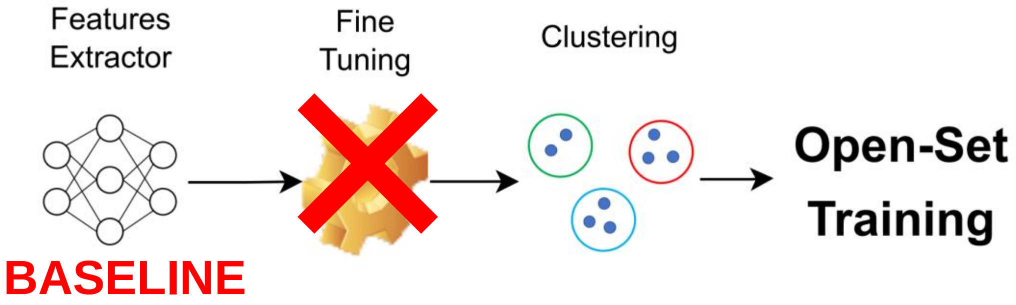
Open-Set training and validation

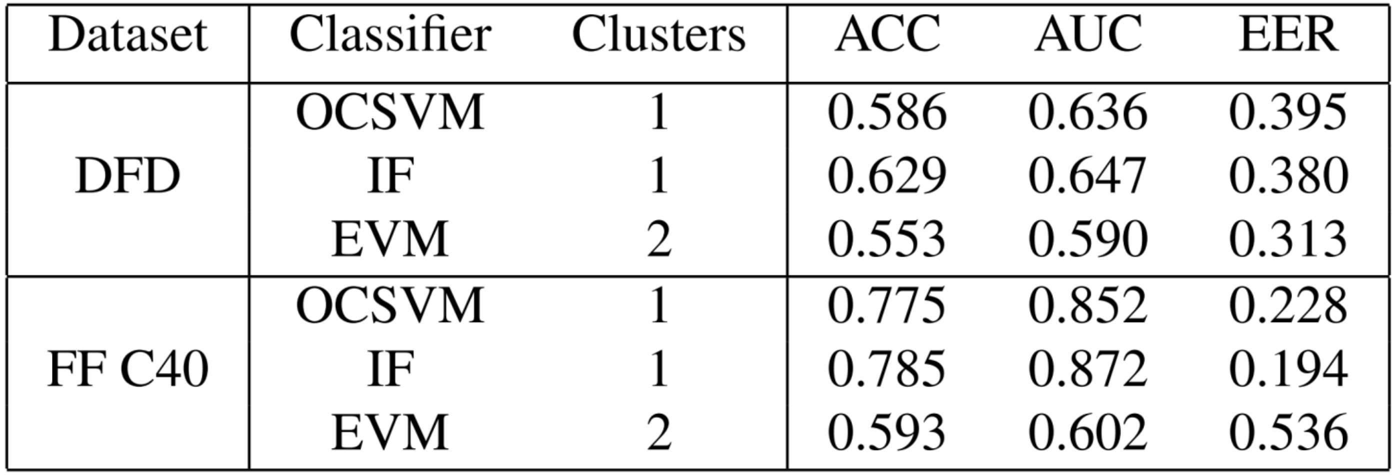
Only Real images from FaceForensics ++ C40

Only Real images from DFD – DeepFake Detection (From Google & Jigsaw)

EXPERIMENTS AND RESULTS

1 – Baseline feature extractor and no Fine Tuning

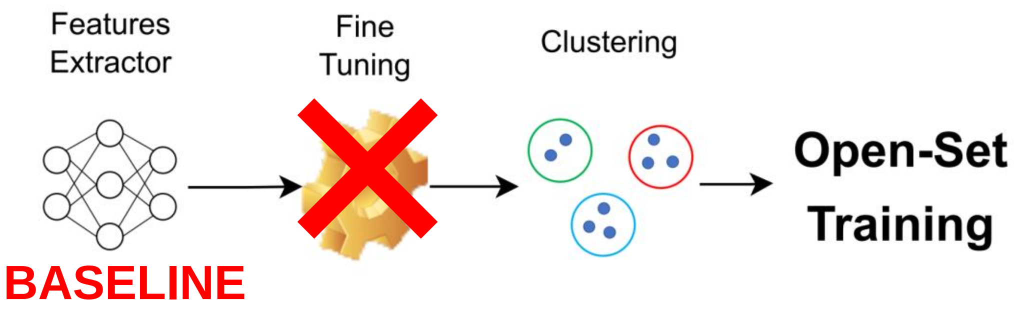




Dataset	Classifier	Clusters	ACC	AUC	EER
DFD	OCSVM	1	0.586	0.636	0.395
	IF	1	0.629	0.647	0.380
	EVM	2	0.553	0.590	0.313
FF C40	OCSVM	1	0.775	0.852	0.228
	IF	1	0.785	0.872	0.194
	EVM	2	0.593	0.602	0.536

EXPERIMENTS AND RESULTS

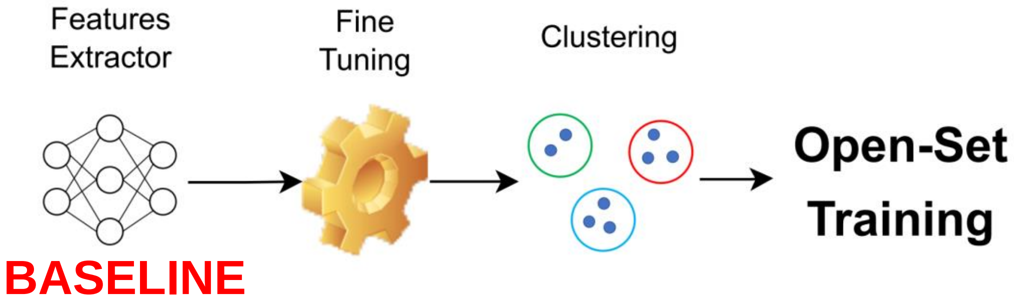
2 – Baseline feature extractor, Dimensionality reduction and no Fine Tuning



Dataset	Classifier	dim	clusters	ACC	AUC	EER
DFD	OCSVM	256	2	0.544	0.580	0.445
	IF	128	3	0.542	0.561	0.449
	EVM	96	4	0.500	0.486	0.969
FF C40	OCSVM	16	3	0.560	0.750	0.068
	IF	32	3	0.559	0.689	0.048
	EVM	32	3	0.631	0.675	0.028

EXPERIMENTS AND RESULTS

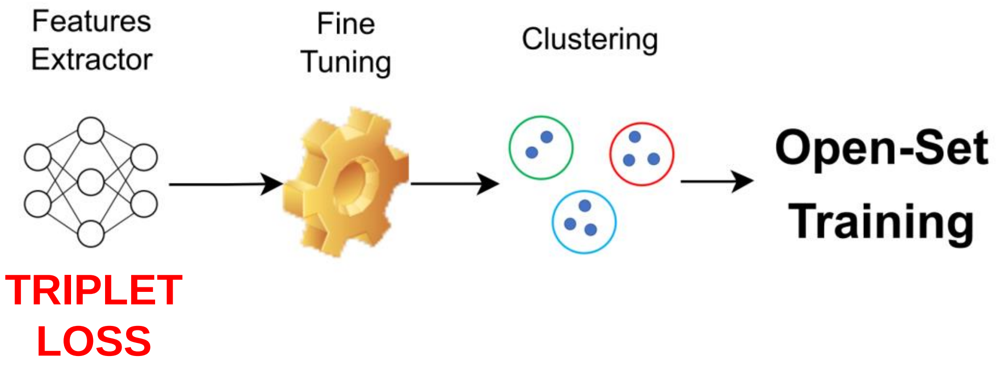
3 – Baseline feature extractor, no dimensionality reduction and Fine Tuning

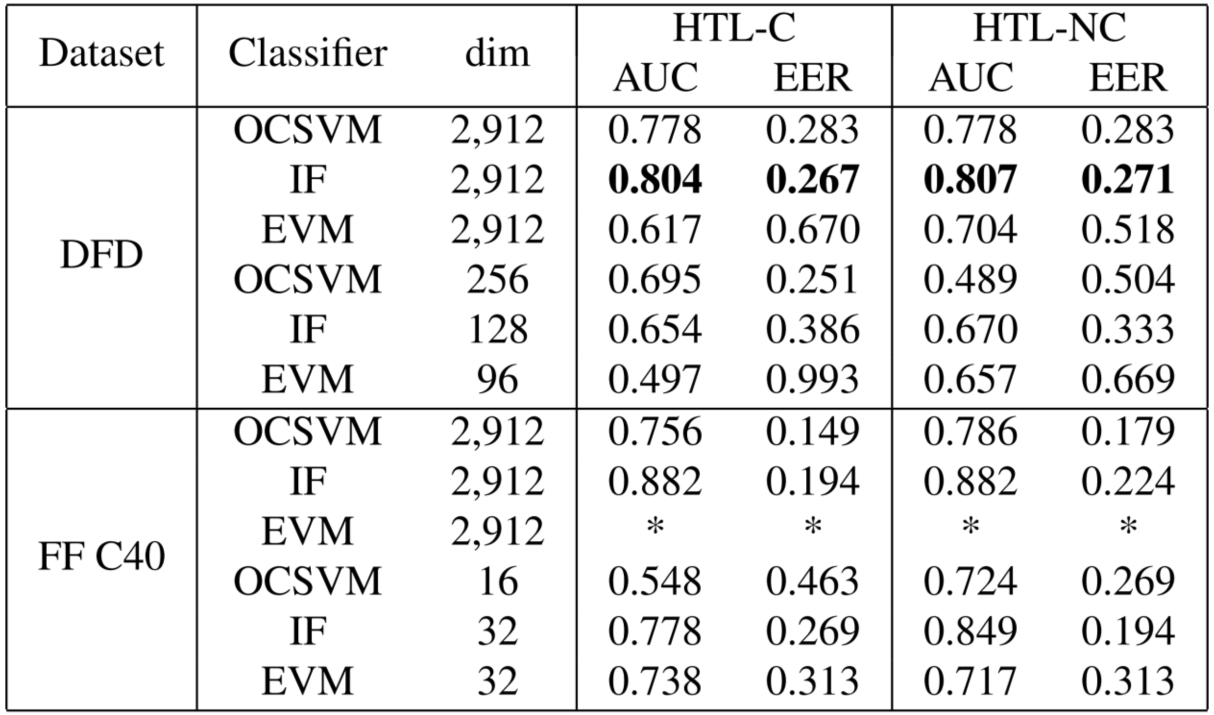


Dataset	Classifier	ACC	AUC	EER
DFD	OCSVM	0.554	0.554	0.559
	IF	0.601	0.630	0.402
	EVM	0.450	0.430	0.800
FF C40	OCSVM	0.764	0.861	0.193
	IF	0.788	0.865	0.207
	EVM	0.531	0.558	0.901

EXPERIMENTS AND RESULTS

4 – Feature extractor with Triplet Loss, Dimensionality reduction and no Fine Tuning





Dataset	Classifier	dim	HTL-C		HTL-NC	
			AUC	EER	AUC	EER
DFD	OCSVM	2,912	0.778	0.283	0.778	0.283
	IF	2,912	**0.804**	**0.267**	**0.807**	**0.271**
	EVM	2,912	0.617	0.670	0.704	0.518
	OCSVM	256	0.695	0.251	0.489	0.504
	IF	128	0.654	0.386	0.670	0.333
	EVM	96	0.497	0.993	0.657	0.669
FF C40	OCSVM	2,912	0.756	0.149	0.786	0.179
	IF	2,912	0.882	0.194	0.882	0.224
	EVM	2,912	*	*	*	*
	OCSVM	16	0.548	0.463	0.724	0.269
	IF	32	0.778	0.269	0.849	0.194
	EVM	32	0.738	0.313	0.717	0.313

Baseline in closed set scenario

- DFD: AUC of 76.23 and an EER of 0.303
- FF C40: AUC of 99.46 and an EER of 0.29

CONCLUSION

- ✓ The open-set approach to deepfake detection is more challenging, but it provides a more robust model against variations in the generation technique.

- ✓ By employing Triplet Loss with Hard mining during feature extractor training, we achieved better results than those obtained with the closed-set approach.

- ✓ Dimensionality reduction and fine-tuning did not yield benefits for our model.

- ✓ The proposed organizational chart can be evaluated using other methods of extraction, dimensionality reduction, clustering, and fine-tuning.

- ✓ Initializing feature extractor weights using self-supervised methods.

UNICAMP

INSTITUTE OF
COMPUTING

THANKS

recod.ai

FEDERAL
INSTITUTE OF
SÃO PAUO

- michael.diniz@ifsp.edu.br

OPEN TO QUESTIONS!