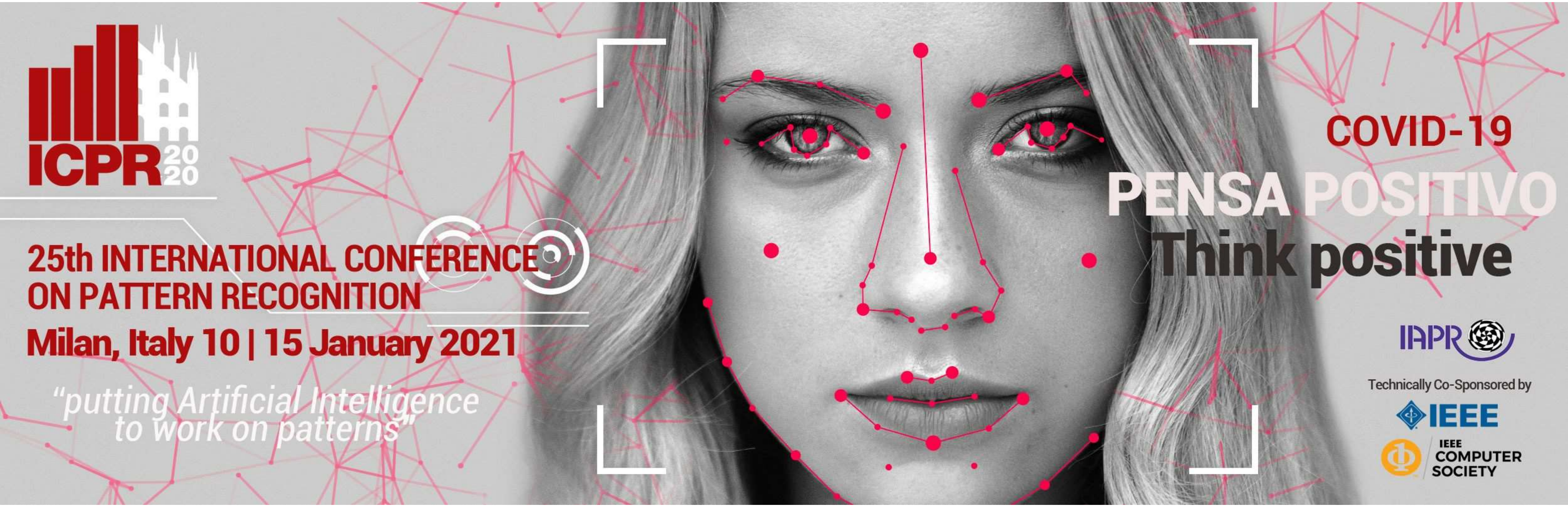


Video Face Manipulation Detection Through Ensemble of CNNs

Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, Stefano Tubaro

ISPL Lab - Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano

POLITECNICO MILANO 1863

ISPL Image and Sound Processing Lab

Motivation

Automatic video editing techniques have made great steps forward in the recent years

Among them **video face manipulation** is one of the most popular

Their **diffusion on the Internet** and social media, and the **availability** to a wide audience of **computer softwares**, source codes and even smartphone apps for generating them are growing concerns for the forensics community

Examples of malicious use: revenge porn, fake news spreading, etc…

Motivation

Face manipulation traces are **subtle** and hard to detect from a signal processing perspective

Different techniques leave different traces with semantically similar results

State of the art for detection: data-driven solution, often based on convolutional neural networks (CNN)

Problem:

- **Difficulties** in generalizing on different manipulation techniques
- **Lack** of insight on what triggered the CNN decision

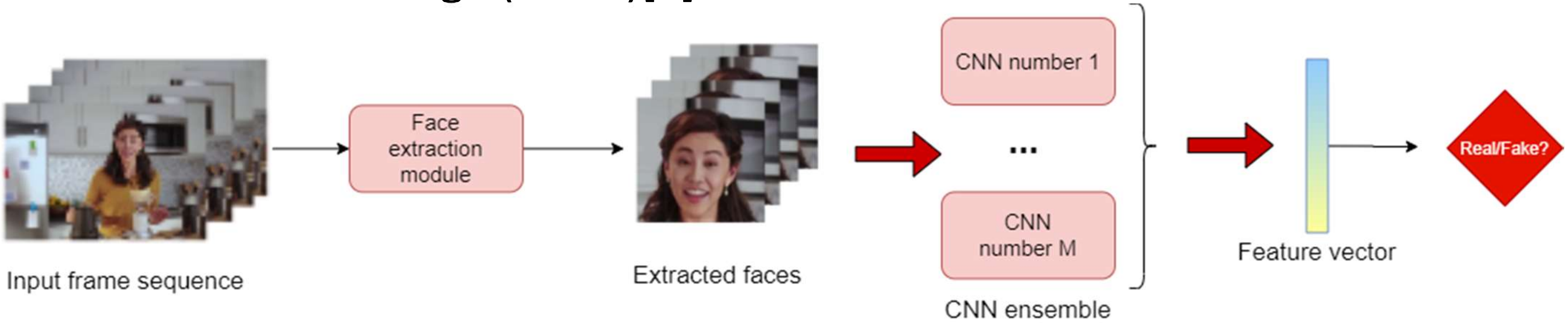
Contribution

Ensemble of CNNs for video face manipulation detection, exploiting:

1. Explicit **attention** mechanisms
2. Different **learning strategies** (**Siamese** + **end-to-end** paradigms)

Goals:

1. **combine** complementary **high-level information** extracted by CNN-based classifiers
2. Develop a lightweight solution => respect the constraint of the Deep Fake Detection Challenge (DFDC)[*]



[*] "DeepFake Detection Challenge Results," https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai.

Experimental setup

Single network: EfficientNetB4 (good tradeoff dimensions/runtime/performances)

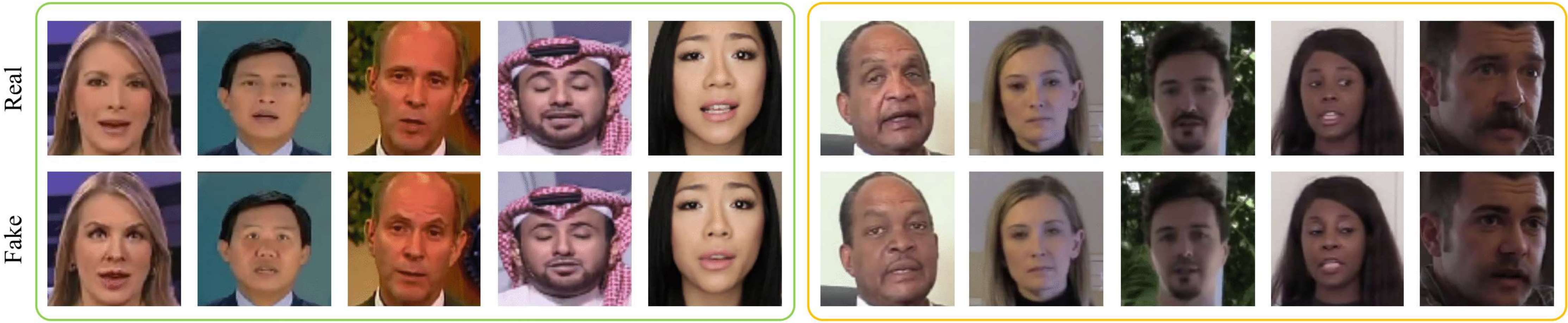
Two datasets:

1. **FaceForensics++ (FF++)**: 4 different face manipulation techniques (2 computer graphics [Face2Face, FaceSwap], 2 learning-based [DeepFakes, NeuralTextures]), 1000 real videos, 4000 fake
2. **Facebook/Kaggle Deep Fake Detection Challenge dataset (DFDC): training** set of the homonymous challenge, made of almost 120000 videos (20000 real, 100000 fake), 8 different manipulation techniques

Train/validation/test split at video level for each dataset

Experimental setup

Datasets samples



FF++ DFDC

Experimental setup

Detection on a frame-per-frame basis, extracting faces from each frame in a pre-processing step:

- Faces are extracted from 32 frames uniformly sampled over time of each video of each dataset
- Faces are cropped with a fixed aspect ratio of 1:1, then resized to a fixed size of 256×256 pixels

The network predicts the likelihood of each face being manipulated

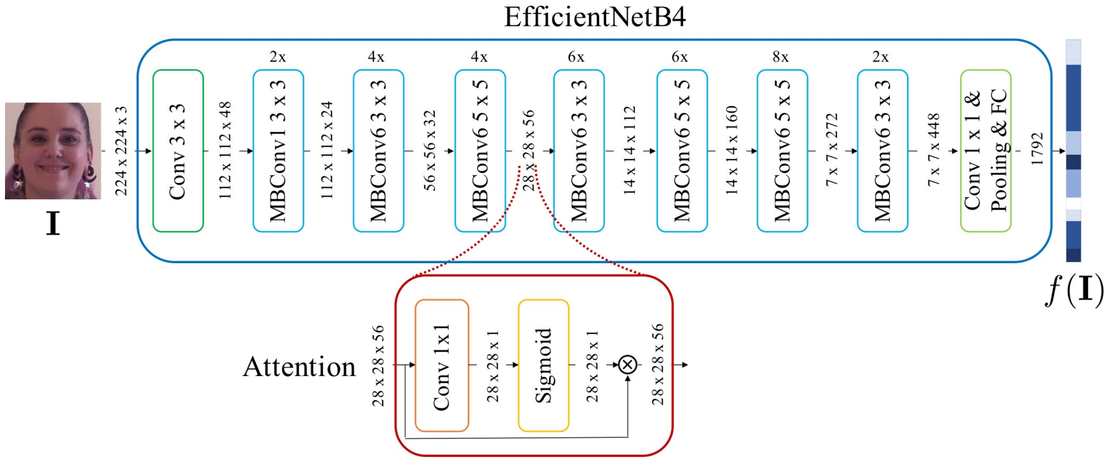
Results reported at frame level:

1. **Area Under the Curve** (**AUC**) of a Receiver-Operating-Characteristic (ROC) curve
2. **Log-loss** values

Case study 1) attention mechanism

Explicit attention mechanism:

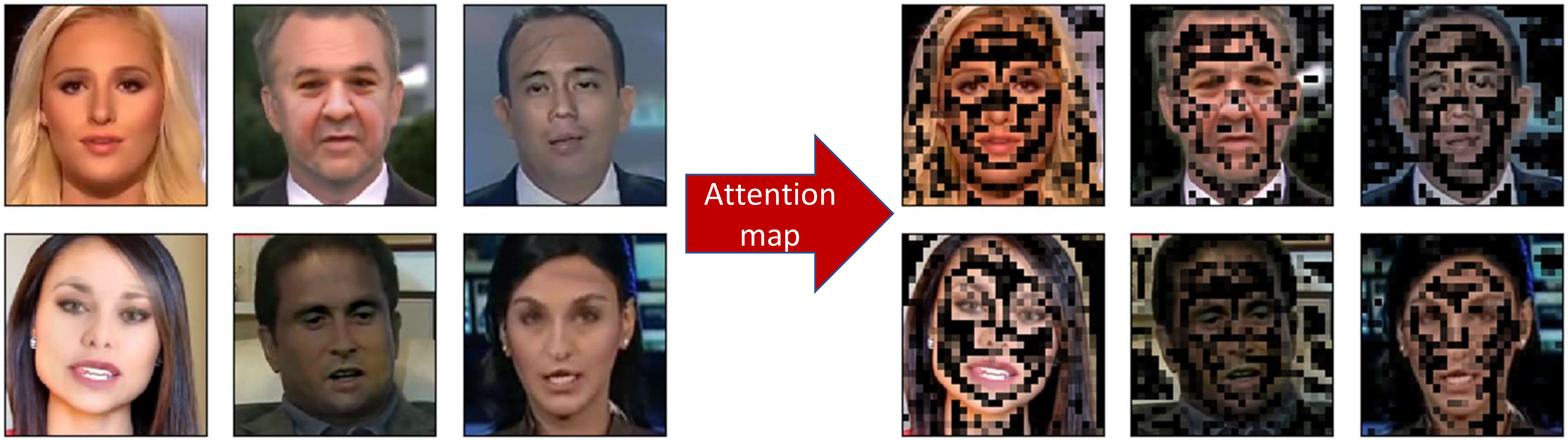
1. Take the feature maps produced by the 3rd MBConvLayer;
2. Process with 1x1 convolution followed by a sigmoid activation;
3. Multiply the resulting attention-map by each feature map of the layer.



Results case study 1) attention mechanism

Explicit attention mechanism

Easily **mapped** on the input **highlights** the **elements** the network judged **more informative**



Case study 2) training paradigm

1. **End-to-end:** feed the network with a face, output a manipulation score; train using a simple **binary-cross entropy (BCE)** loss
2. **Siamese:**
    1. train the network first as a feature extractor using the **triplet margin loss[*];**
    2. finetune the network using BCE;

Idea: develop a feature descriptor from data that privileges similarities of samples belonging to the same class

Goal: obtain a representation in the network's encoding space separating nicely manipulated and non-manipulated samples

[*] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking"

Results case study 2) training paradigm

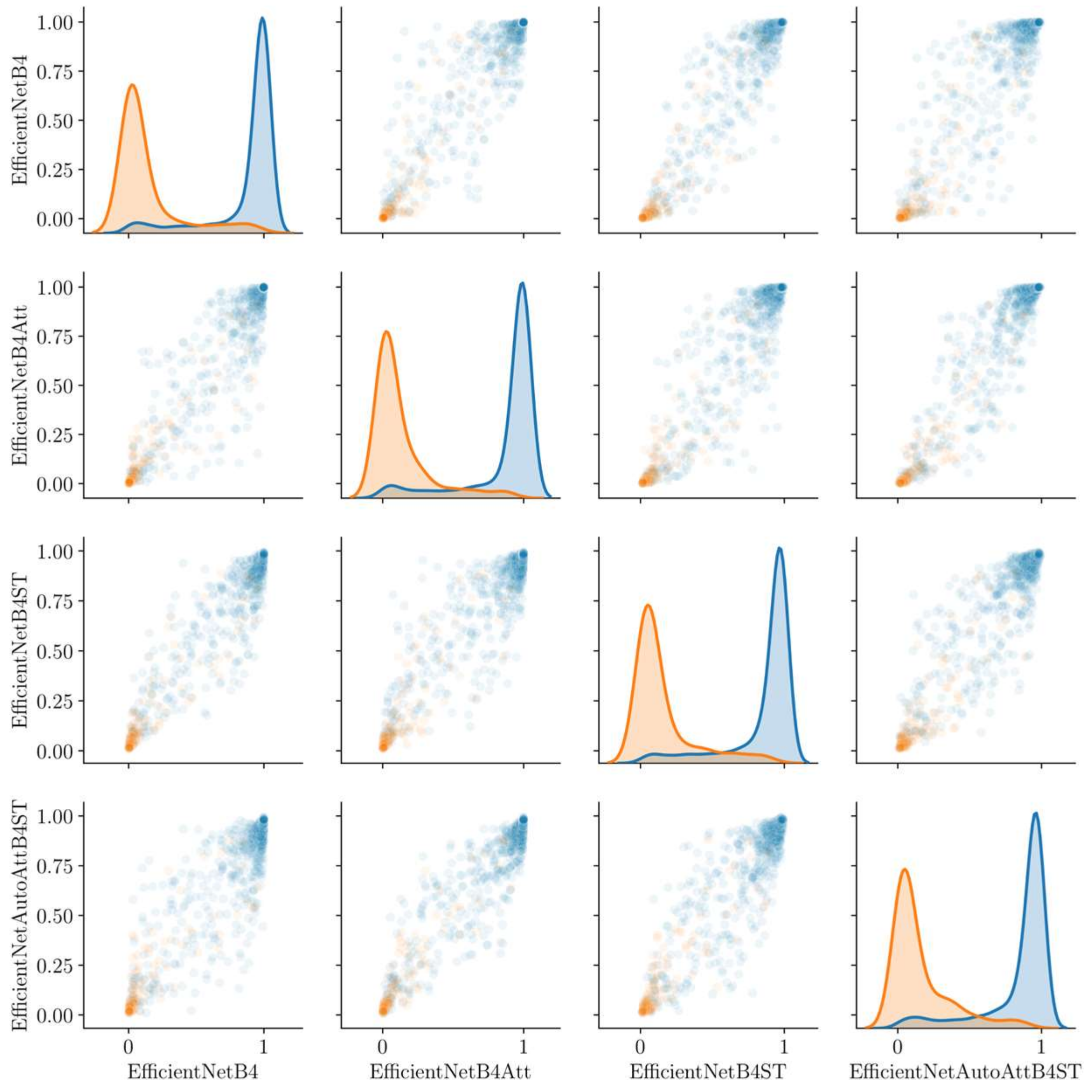
- t-SNE projection of 20 FF++ videos
- Each point = **feature** descriptor of a video **frame face** extracted by a CNN trained with the **siamese** approach
- 32 frames per video

Case study 3) networks independence

1. **Train** 4 different networks on FF++ and DFDC train splits separately:
    1. **EfficientNetB4:** no attention, end-to-end paradigm;
    2. **EfficientNetB4Att:** attention mechanism, end-to-end paradigm;
    3. **EfficientNetB4ST:** no attention, Siamese paradigm;
    4. **EfficientNetB4AttST:** attention mechanism, Siamese paradigm.
2. **Test** on the test split of each dataset separately
3. **Look** at the **distributions** of the **manipulation scores** of the test samples

Results case study 3) networks independence

- Detection scores on DFDC
- Each network returns a slightly different score for each frame
- An ensemble may benefit from different "perspectives" given by each network



Experimental results: detection capability

Combine the networks seen previously (combination of 2, 3 and 4 networks)

Compare the performances against a XceptionNet (baseline used for FF++[*])

Train and **test** again separately on the two datasets

Ensembles provide always **better results** both in accuracy (AUC) and quality (Log-Loss) of the detection

TOP-3 RESULTS PER COLUMN IN **BOLD**, BASELINE IN *ITALICS*

Xception Net	EfficientNet				AUC		LogLoss	
	B4	B4ST	B4Att	B4AttST	FF++	DFDC	FF++	DFDC
✓					*0.9273*	*0.8784*	*0.3844*	*0.4897*
	✓				0.9382	0.8766	0.3777	0.4819
		✓			0.9337	0.8658	0.3439	0.5075
			✓		0.9360	0.8642	0.3873	0.5133
				✓	0.9293	0.8360	0.3597	0.5507
	✓	✓			0.9413	**0.8800**	0.3411	0.4687
	✓		✓		0.9428	**0.8785**	0.3566	0.4731
	✓			✓	0.9421	0.8729	0.3370	0.4739
		✓	✓		0.9423	0.8760	0.3371	0.4770
		✓		✓	0.9393	0.8642	**0.3289**	0.4977
			✓	✓	0.9390	0.8625	0.3515	0.4997
	✓	✓	✓		**0.9441**	**0.8813**	0.3371	**0.4640**
	✓	✓		✓	0.9432	0.8769	**0.3269**	**0.4684**
	✓		✓	✓	**0.9433**	0.8751	0.3399	0.4717
		✓	✓	✓	0.9426	0.8719	0.3304	0.4800
	✓	✓	✓	✓	**0.9444**	0.8782	**0.3294**	**0.4658**

[*] A. R¨ossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images"

Experimental results: real-world scenario application

We participated as the **ISPL** team to the DFDC challenge

Hardware and time constraints:

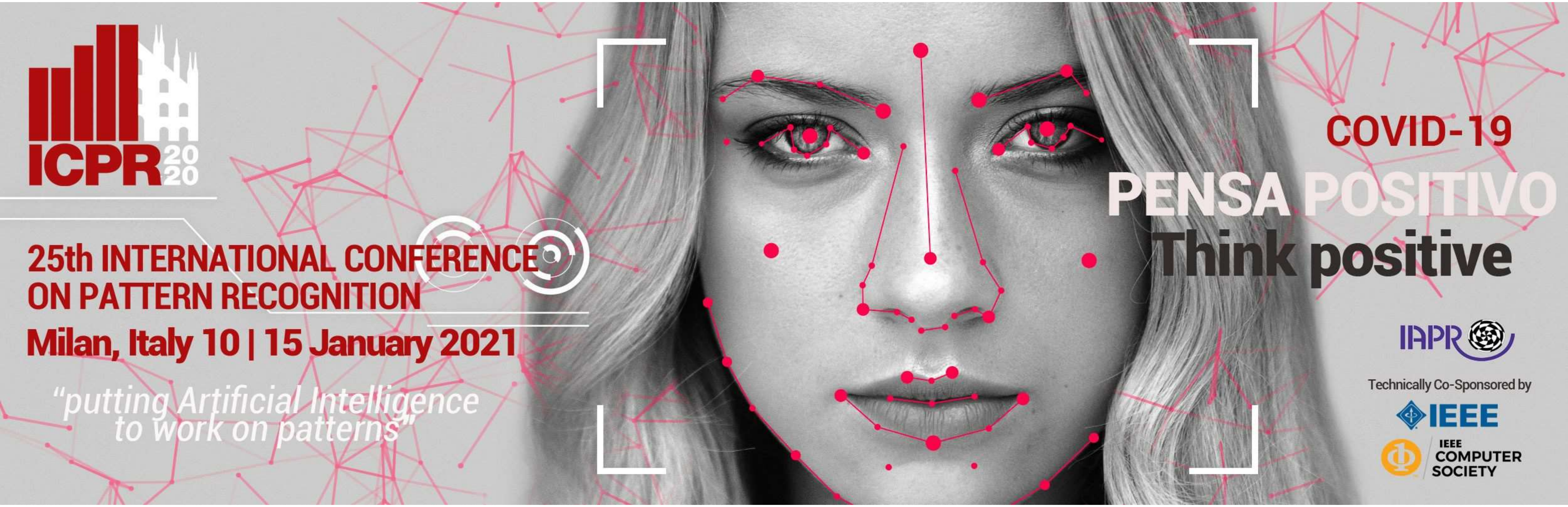
1. The proposed solution had to analyse more than **4000** videos in less than **9 hours** using at most a **single GPU**
2. The proposed solution must not exceed **1GB** of disk space

We participated using an ensemble of the 4 models described previously

Our solution reached **top 2%** (**41st position** among more than 2000 participants) on the final leader board computed against the private test set

Conclusions

1. We developed a CNNs ensemble method for facial manipulation detection
2. We use a single CNN model trained with different paradigms and small architectural variations
3. Our proposed solution, while achieving valid results on two publicly available datasets, is able to provide human comprehensible inference of the model
4. While still offering competitive results, our method keeps computational complexity at bay



Video Face Manipulation Detection Through Ensemble of CNNs

Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, Stefano Tubaro

ISPL Lab - Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano

POLITECNICO MILANO 1863

Thanks for the attention!

ISPL Image and Sound Processing Lab