COMBINING EFFICIENTBET AND VISION TRANSFORMERS FOR VIDEO DEEPFAKE DETECTION

Davide Coccomini, Nicola Messina, Claudio Gennaro, Fabrizio Falchi

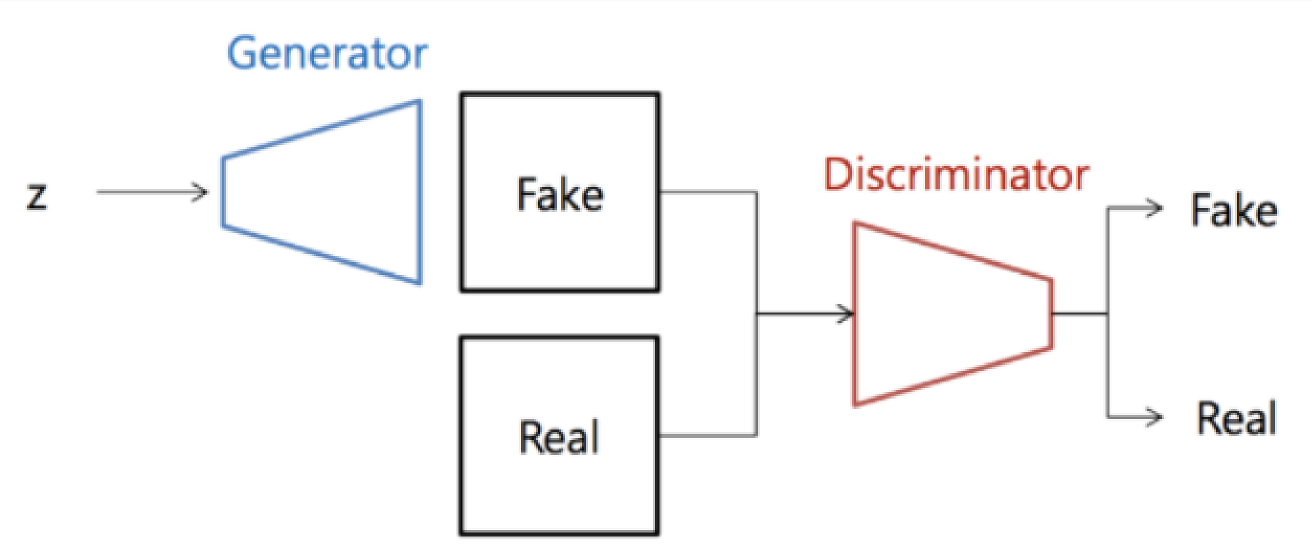
2021.08.05

세종대학교 무인이동체공학과

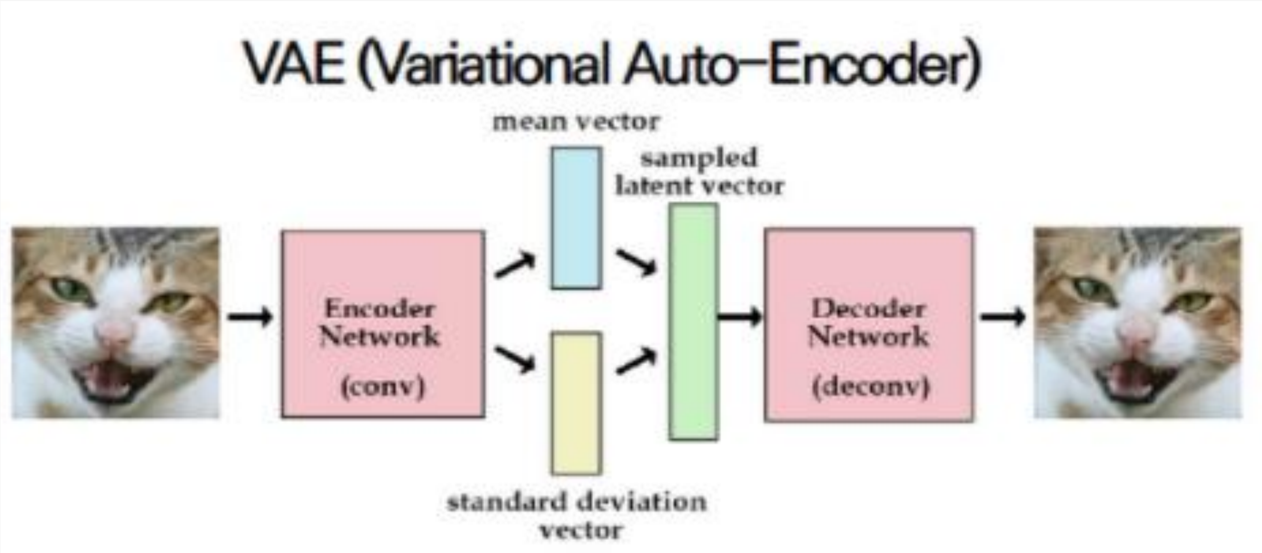
신우정

Deepfake

1. GAN (Generative Adversarial Network)



2. VAE (Variational AutoEncoder)



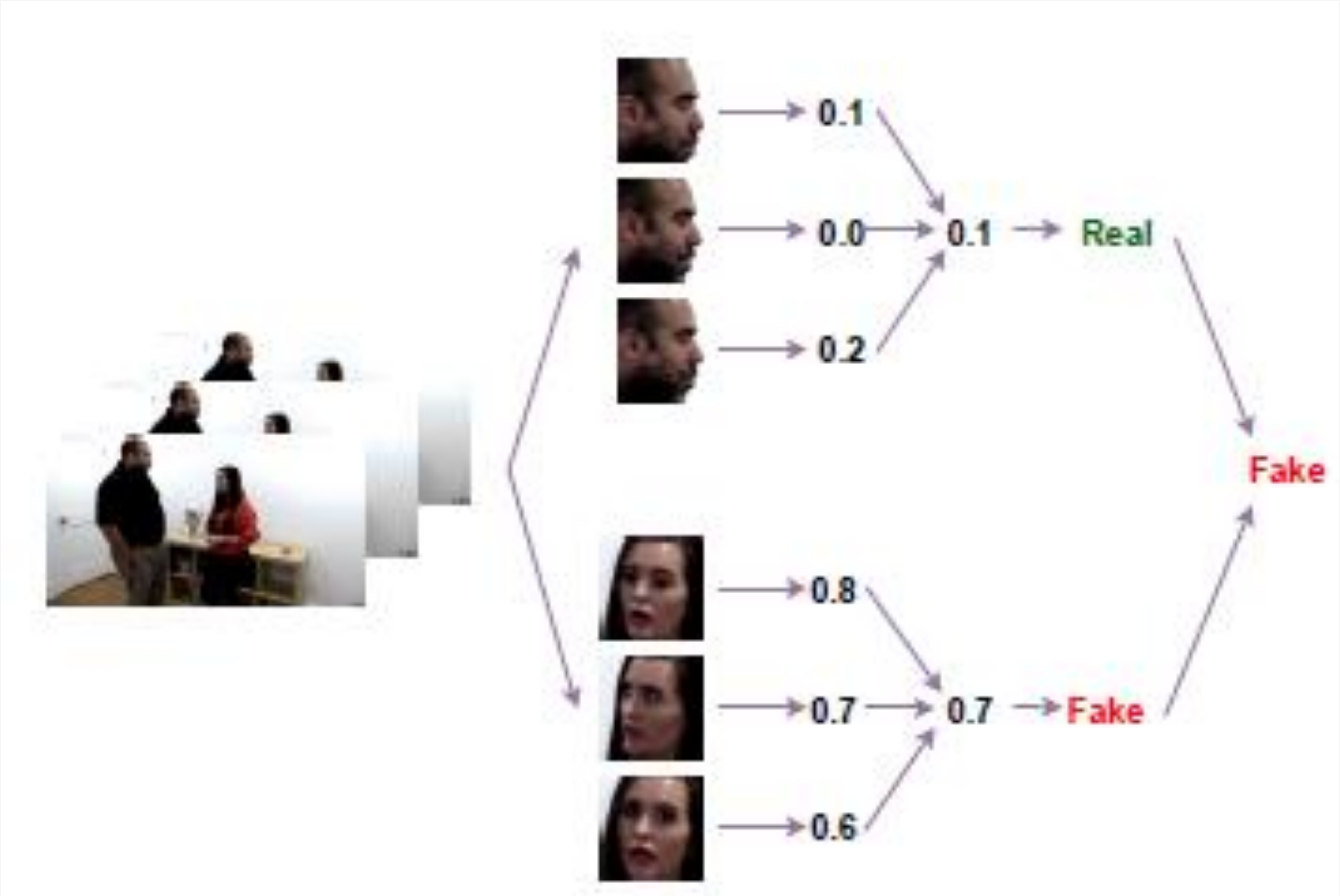
https://www.creativebloq.com/features/deepfake-examples
https://ratsgo.github.io/generative%20model/2017/12/20/gan/
https://velog.io/@ohado/%EB%94%A5%EB%9F%AC%EB%8B%9D-%EA%B0%9C%EB%85%90-1.-VAEVariational-Auto-Encoder

Experiments

<Dataset>
- FaceForensics++
- DFDC dataset

<Training>
- SGD optimizer
- Learning rate -> 0.01

<Inference>
Real/Fake threshold -> 0.55



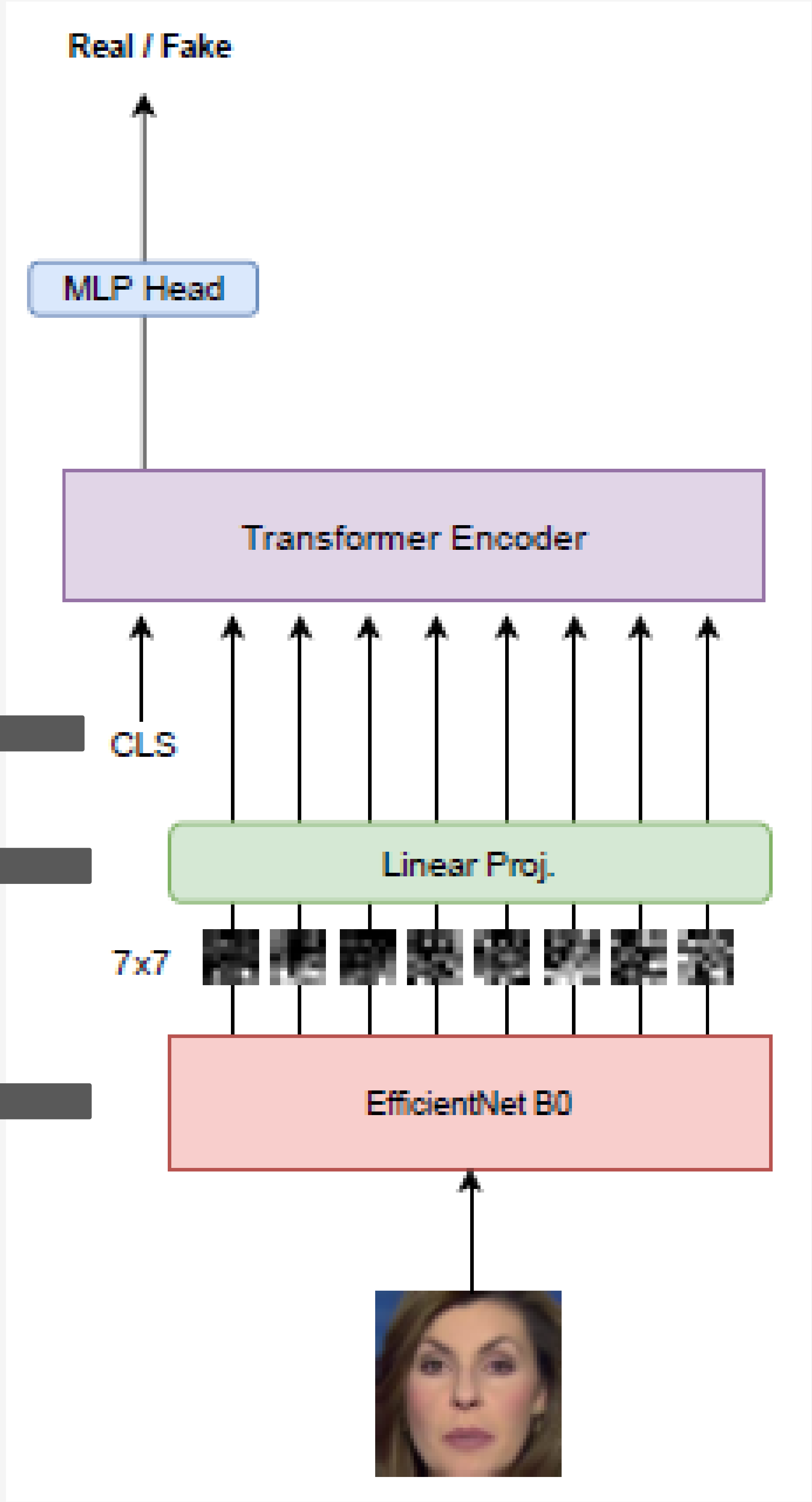
Efficient ViT architecture

결과

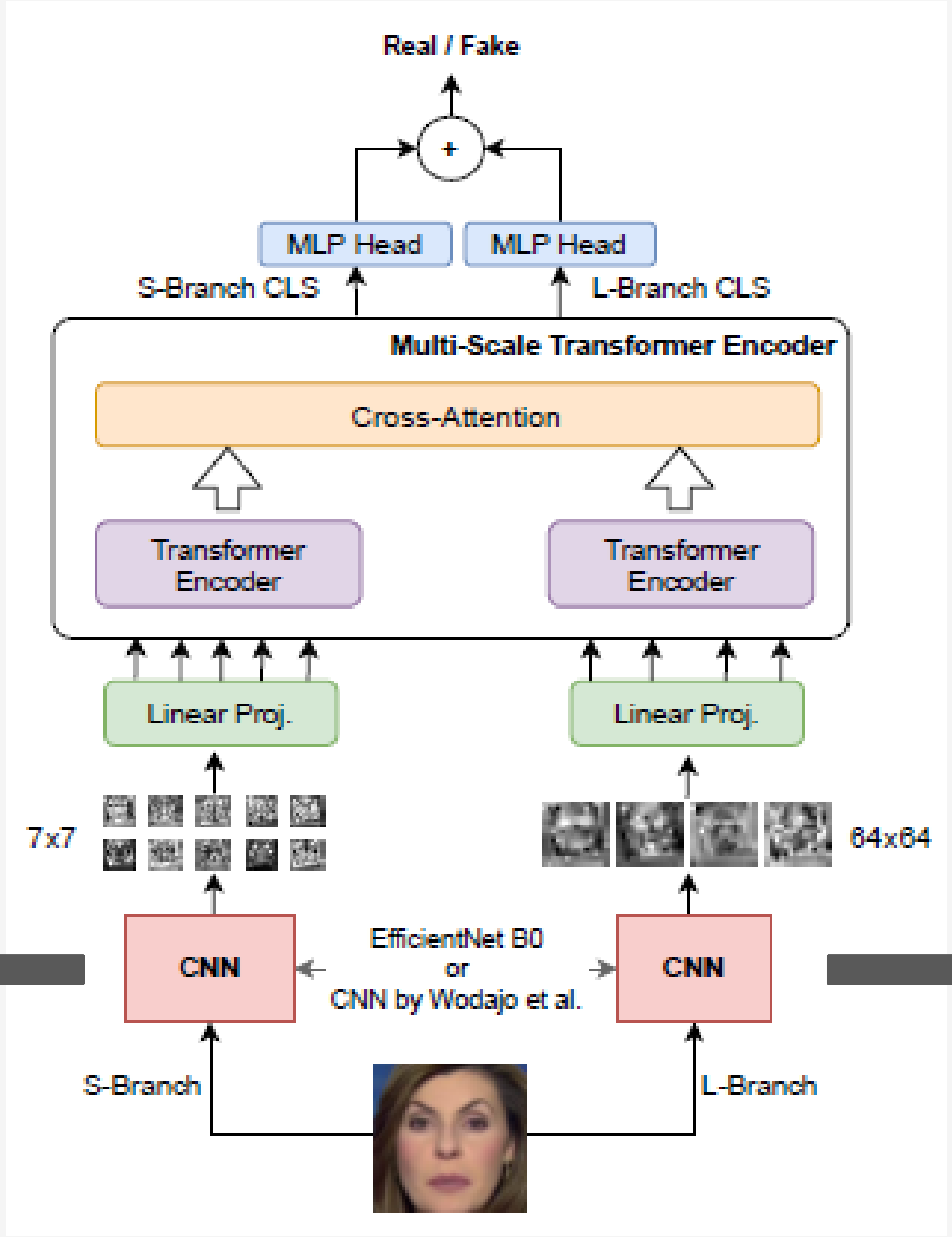
이진 분류 스코어 생성

선형 투영

특징 추출



Convolutional Cross ViT architecture



Small patch
-> local view

Large patch
-> global view

Results

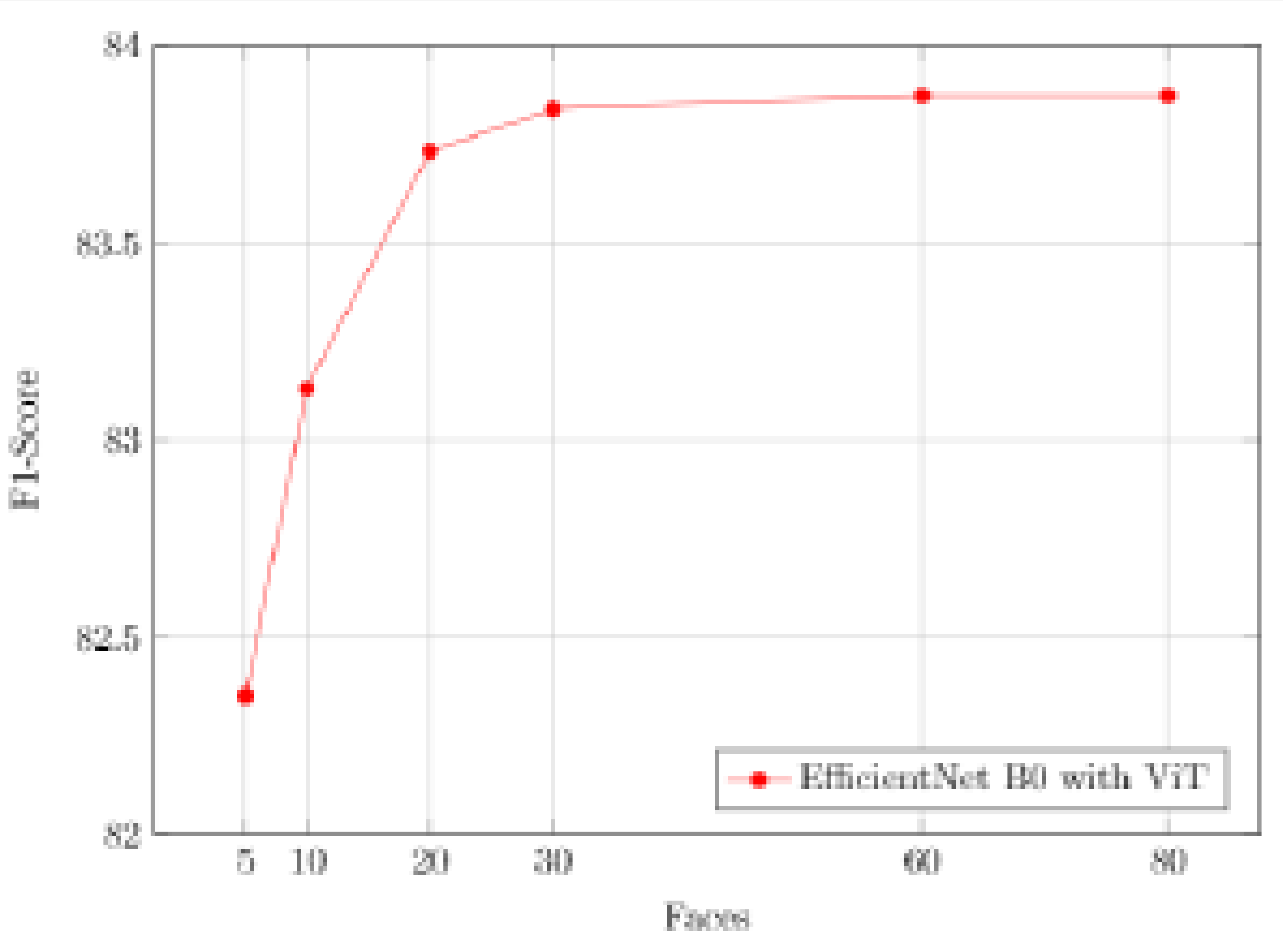
Table 1: Results on DFDC test dataset

Model	AUC	F1-score
ViT with distillation [Heo et al., 2021]	0.978	91.9%
Selim EfficientNet B7 [Seferbekov, 2020]	0.972	90.6%
Convolutional ViT	0.843	77.0%
Efficient ViT (our)	0.919	83.8%
Convolutional Cross ViT (our)	0.925	84.5%
Efficient Cross ViT (our)	0.951	88.0%

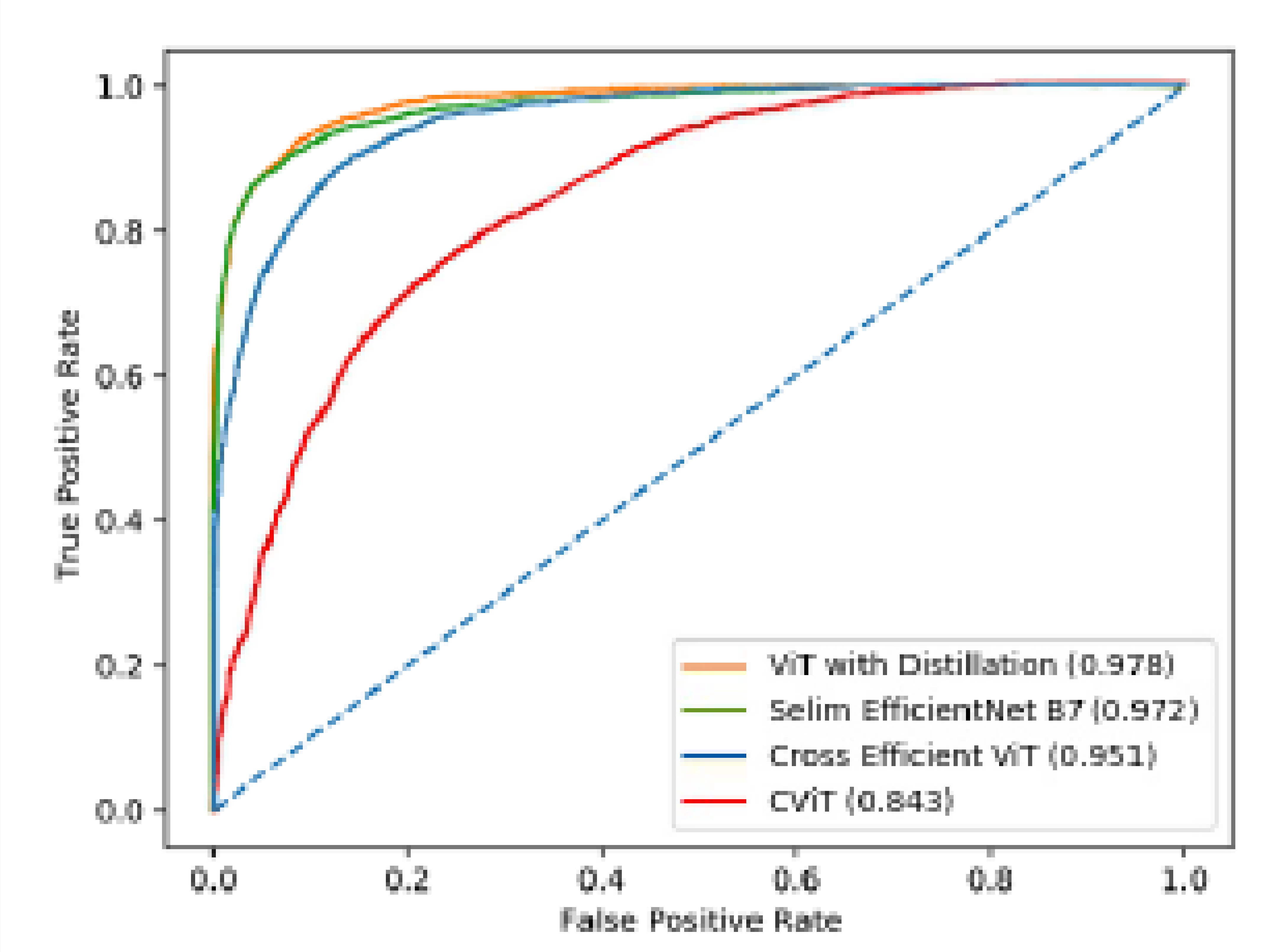
Table 2: Models accuracy on FaceForensics++

Model	Mean	FaceSwap	DeepFakes	FaceShifter	NeuralTextures
Convolutional ViT [Wodajo and Atnafu, 2021]	67%	69%	**93%**	46%	60%
Efficient ViT (our)	76%	78%	83%	76%	68%
Convolutional Cross ViT (our)	76%	81%	83%	73%	67%
Efficient Cross ViT (our)	**80%**	**84%**	87%	**80%**	**69%**

F1-score versus the number of extracted faces



ROC Curves comparison



감사합니다