IMT SCHOOL FOR ADVANCED STUDIES LUCCA

MD2SL

Data Science & Statistical Learning | II Level Master

UNIVERSITÀ DEGLI STUDI FIRENZE

Master Thesis

DeepFake Detection Exploiting Self-Attention Maps

Supervisor: *Prof. Lorenzo Seidenari*

Co-supervisor: *Luca Cultrera, PhD*

Gianmarco Santoro

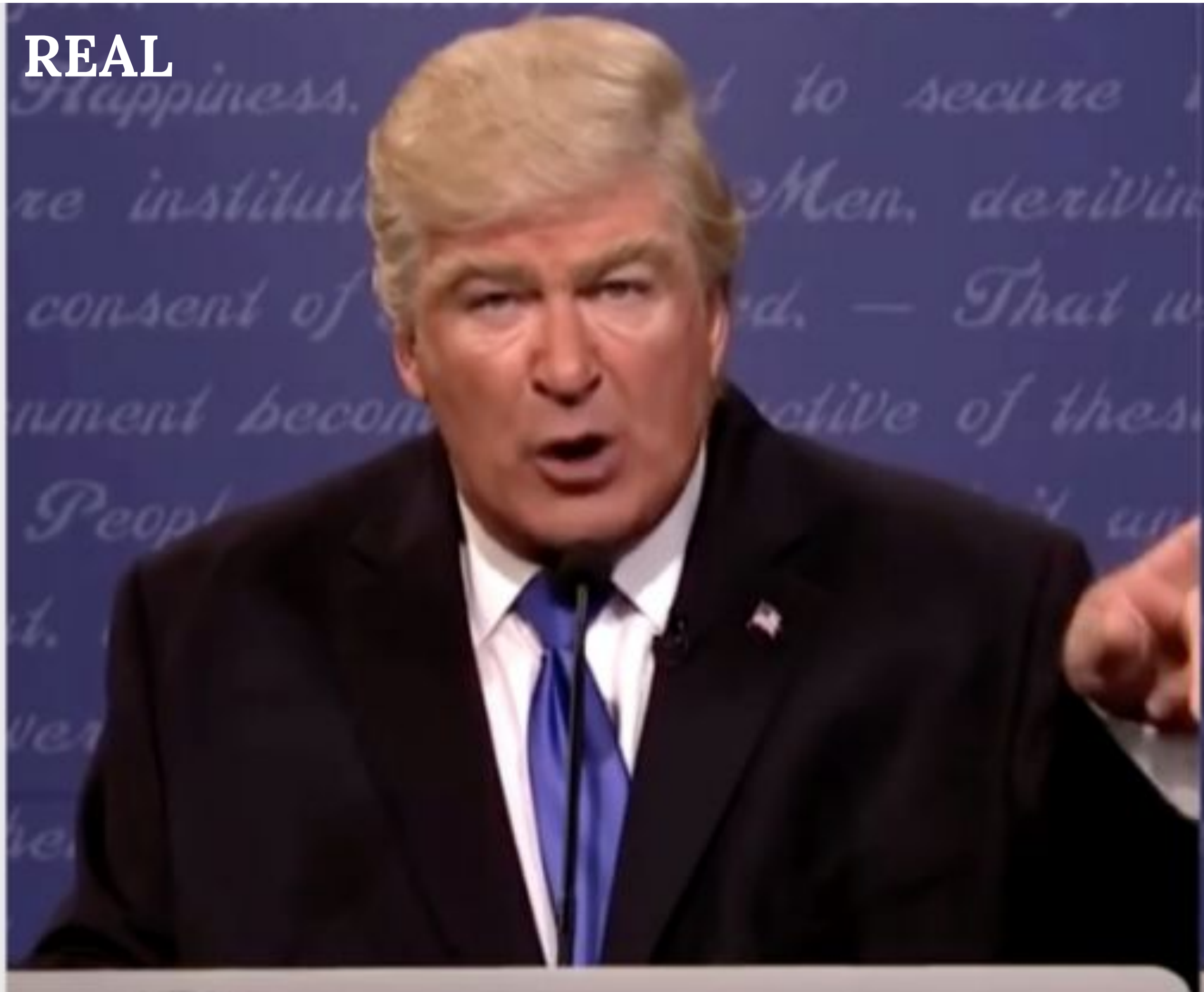
24/04/2024

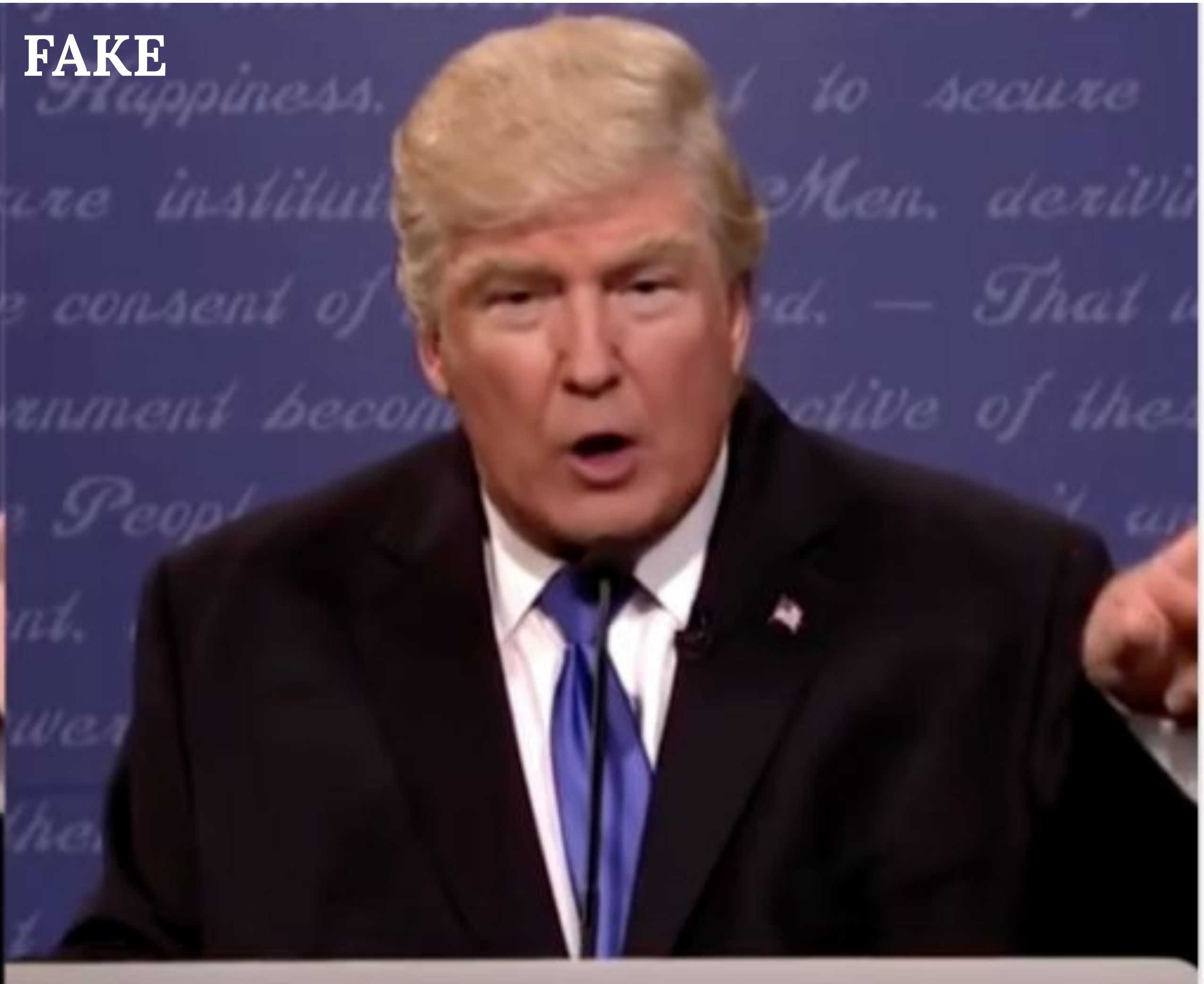
DeepFake Detection | The DeepFake Dilemma

REAL or FAKE?

Benefit → Special effects

Problem → Disinformation





Supervised methods

- Trained on specific forgeries → cannot detect unseen ones
- Very high accuracy in detection

Original

✓

DeepFakes

✓

Face2Face

✓

FaceSwap

✓

FaceShifter

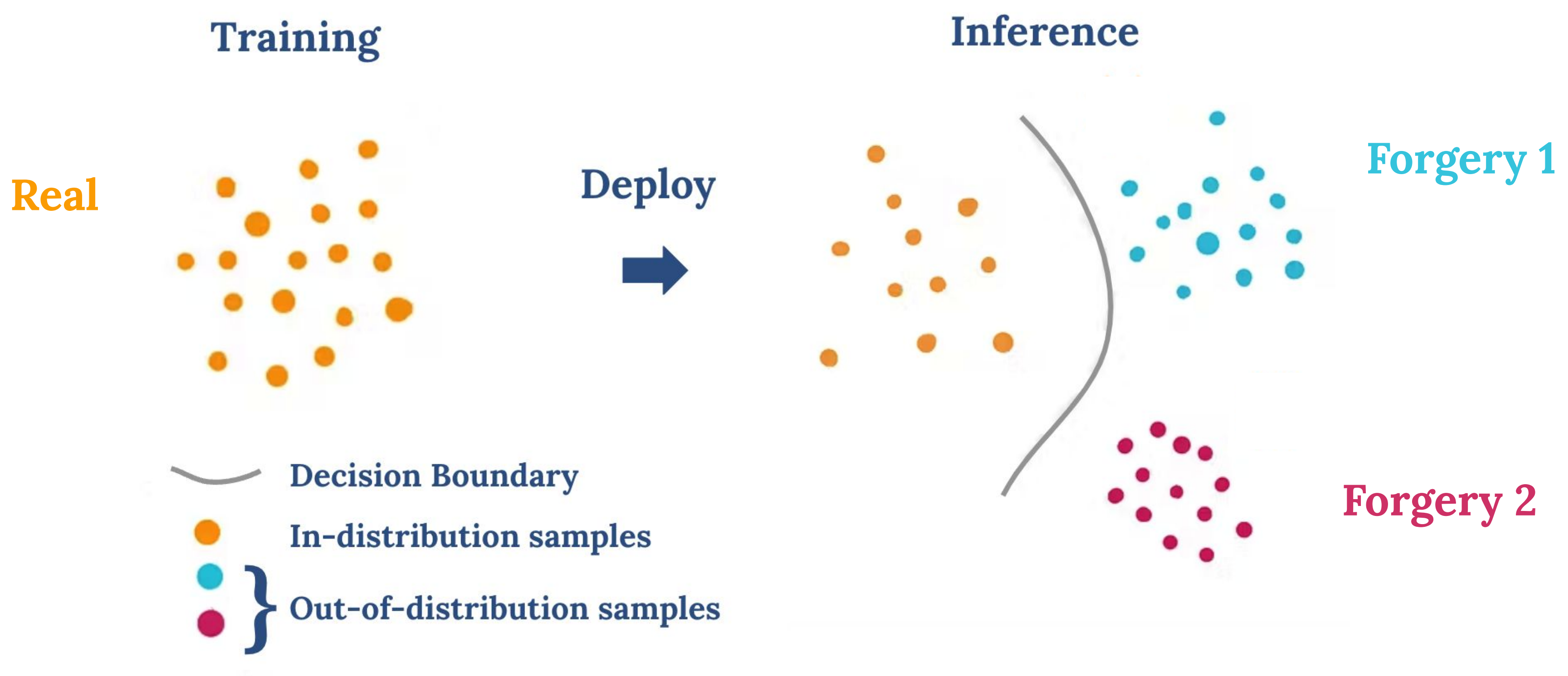
✗

NeuralTextures

✗

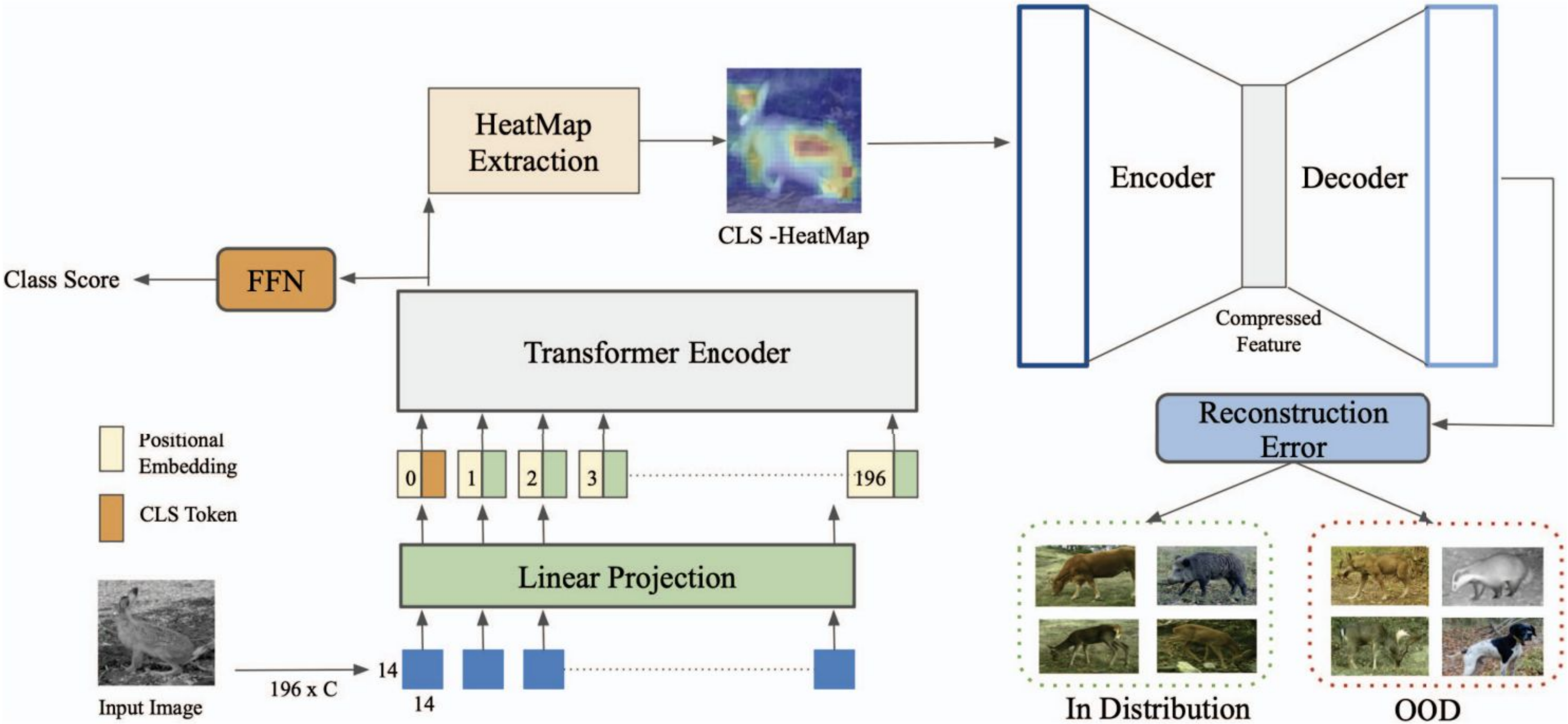
Out-of-Distribution Detection

- Identifying data different from training distribution
- In this case between 2 classes
  - ○ Real images → In-Distribution
  - ○ Fake images → Out-of-Distribution



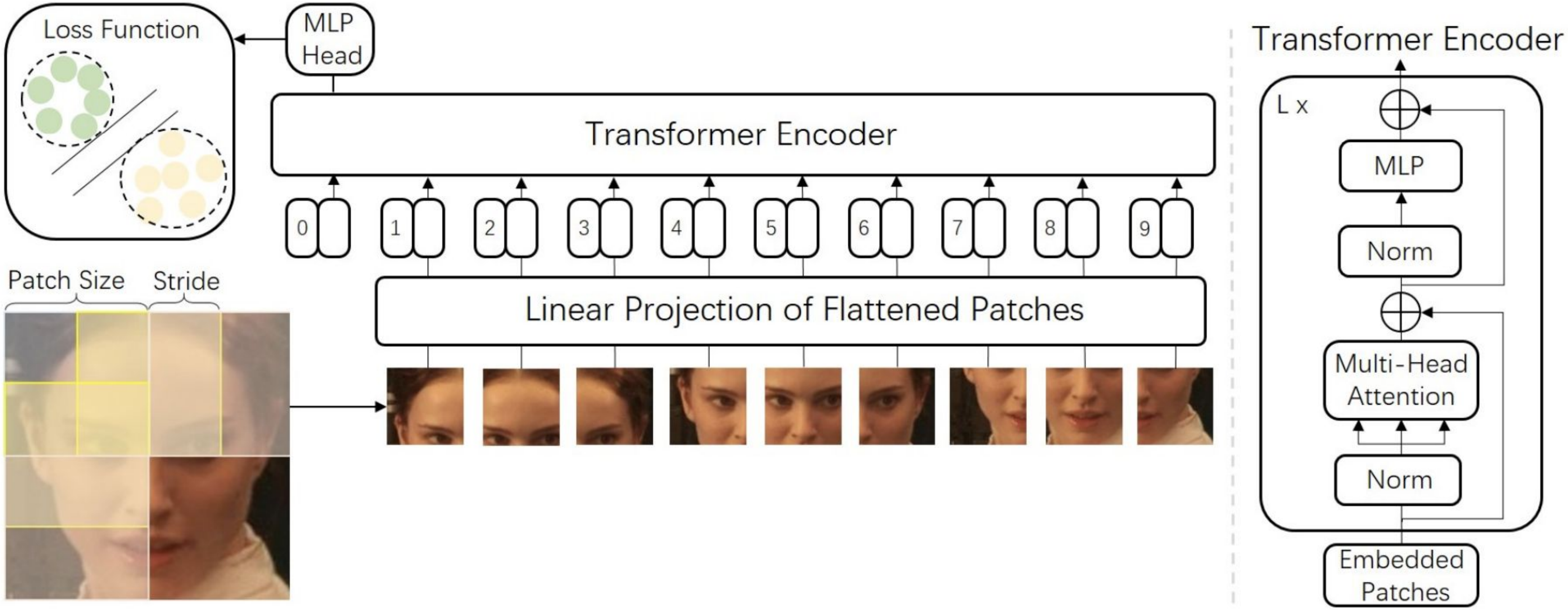
Unsupervised method

1. ViT → extract Attention
2. Conv-AE → discern between In-distribution and Out-of-Distribution images



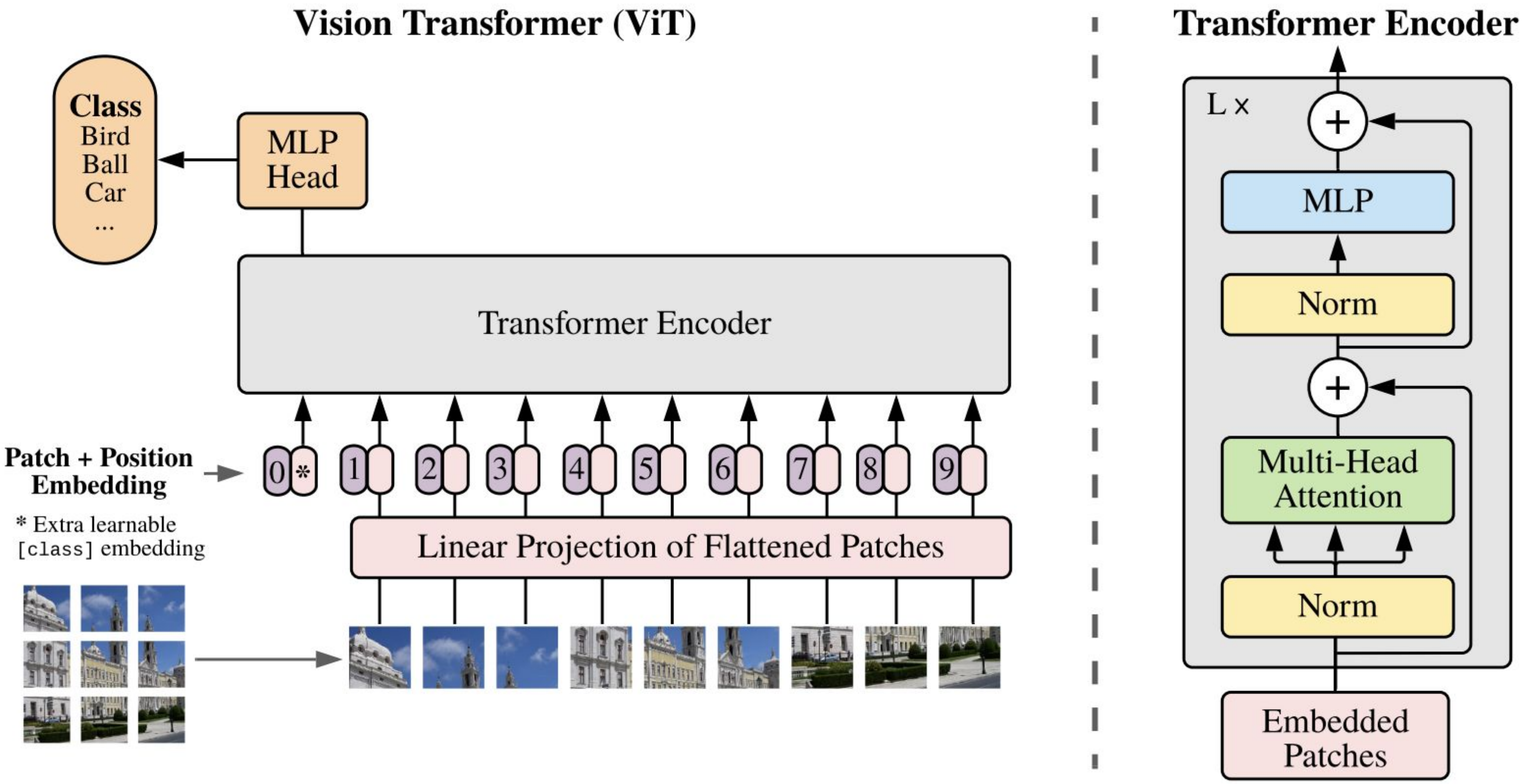
ViT *Face-Transformer*

- Originally → Face recognition: identity
- Out method → Feature extraction: Attention

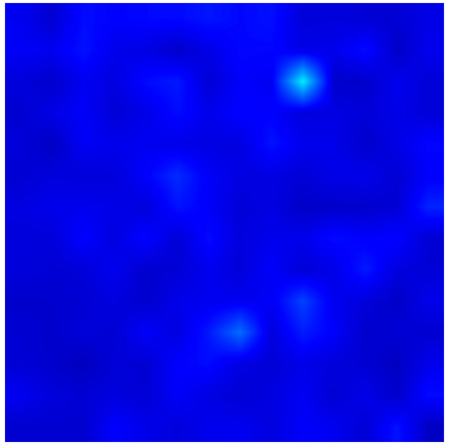


Example - Vision Attention in Vision Transformer

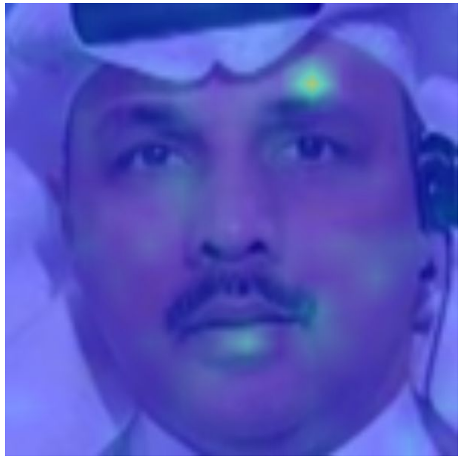
- Architecture → Original ViT
- Vision Attention → from this research



Attention Heatmap

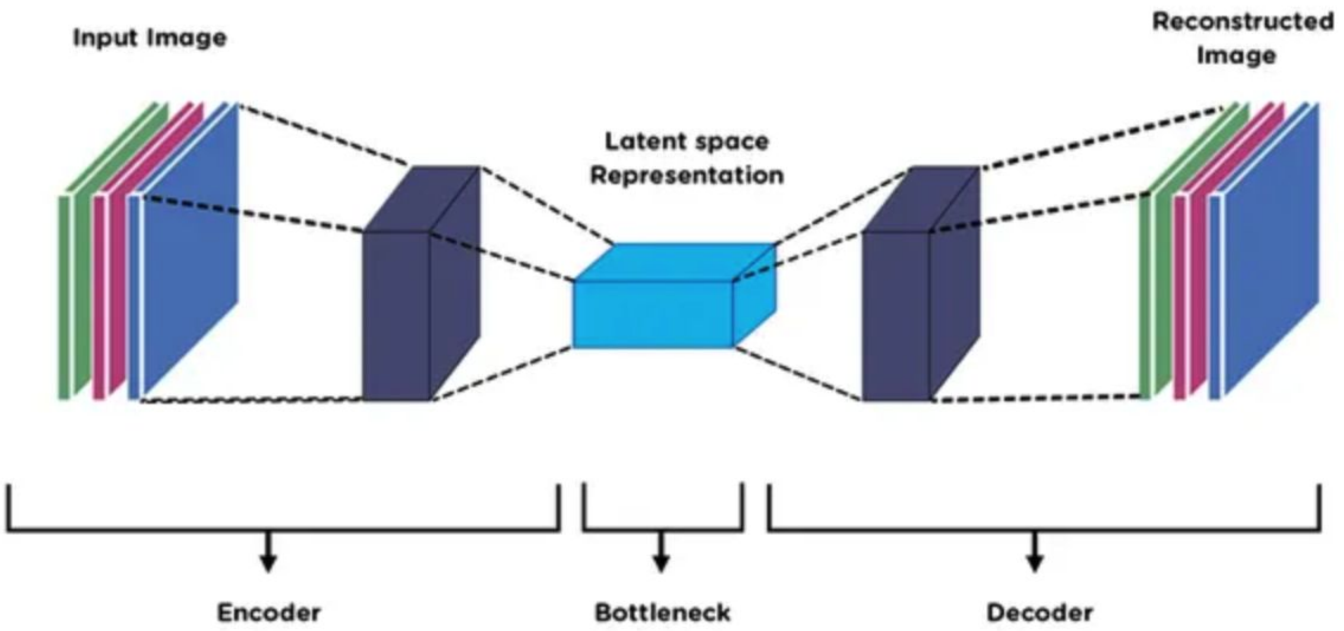


Heatmap on Frame



Conv-AE

- Architecture trained to reconstruct its input accurately
- How:
  - Encoder compresses input in lower-dimensional latent space
  - Decoder reconstruct original input from compressed representation



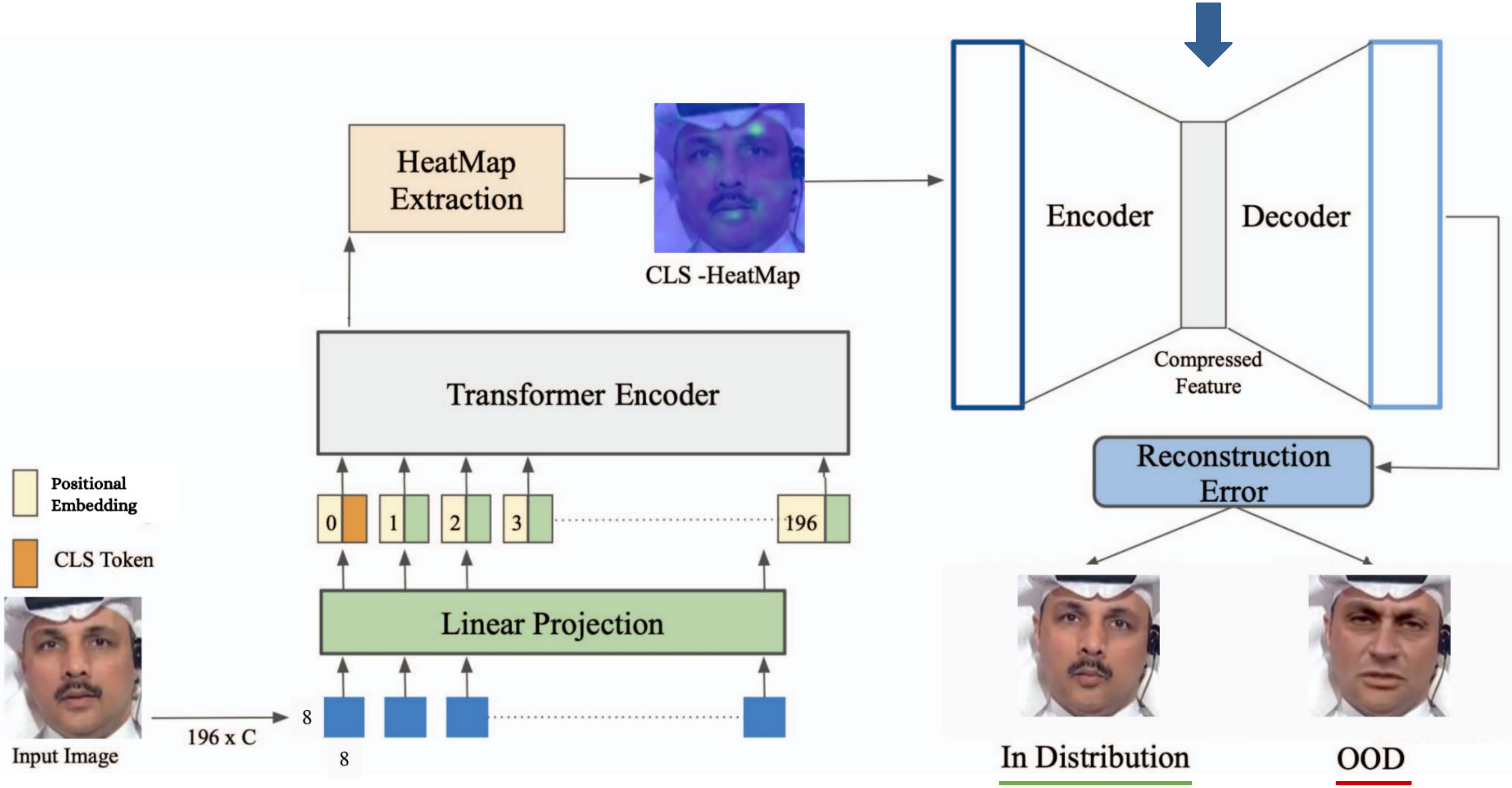
DeepFake Detection | Data Pipeline

Data → ViT → Attention extraction → HeatMaps Dataset →

→ Train Conv-AE on REALs → Test on All Images →

→ Reconstruction Error: Real or Fake

Trained on In-Distribution images only



Test on all images: Real & Fake

Extract Heatmaps of all images: Real & Fake

FaceForensics++ dataset

- 1000 videos from YouTube
  - 1 person per video
- ~ 100 frames per video
- 5 Forgeries + Real

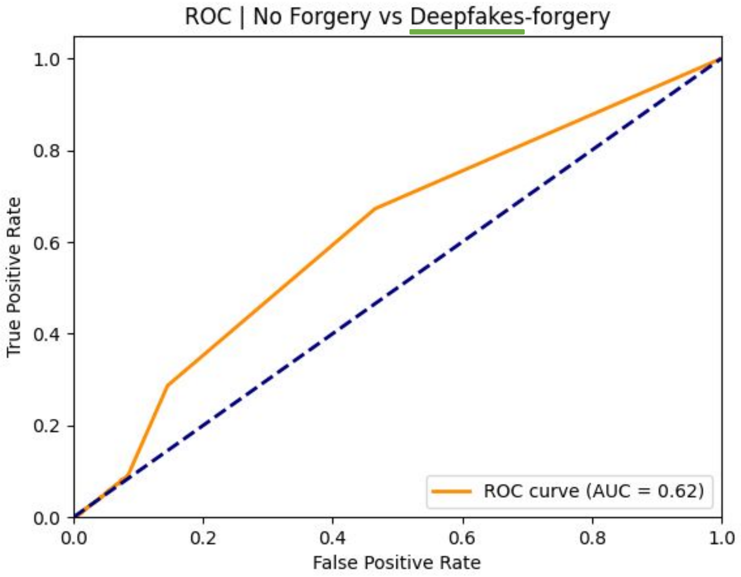
- Dataset split on videos
  - Train → 80 %
  - Validation → 10 %
  - Test → 10 %

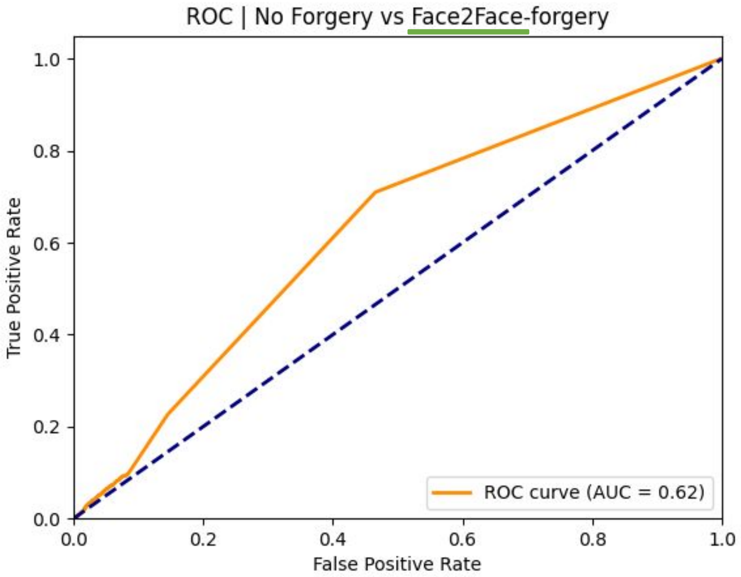
LEGEND:
- IN DISTRIBUTION
- OUT-OF-DISTRIBUTION

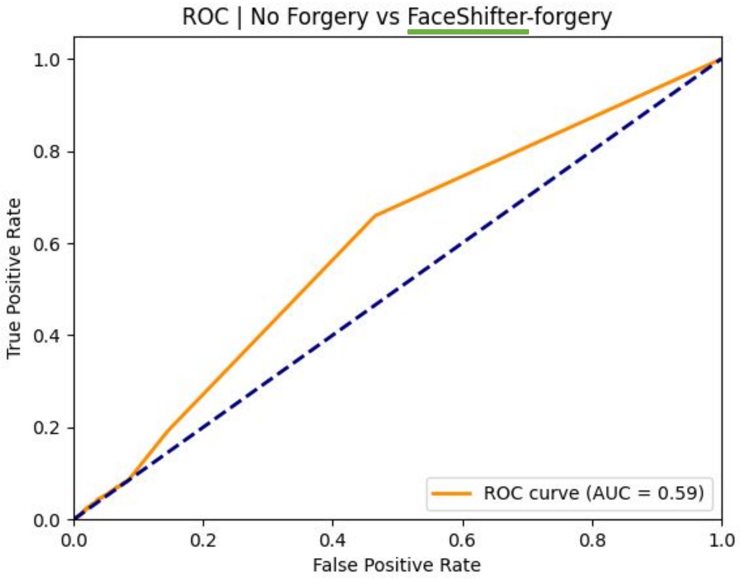
		DATASETS		
		TRAINING	VALIDATION	TEST
IMAGES	REAL	79’954	9'995	10'000
	DEEPFAKES	✖	✖	10'000
	FACE2FACE	✖	✖	10'000
	FACESHIFTER	✖	✖	10'000
	FACESWAP	✖	✖	10'000
	NEURAL TEXTURES	✖	✖	10'000

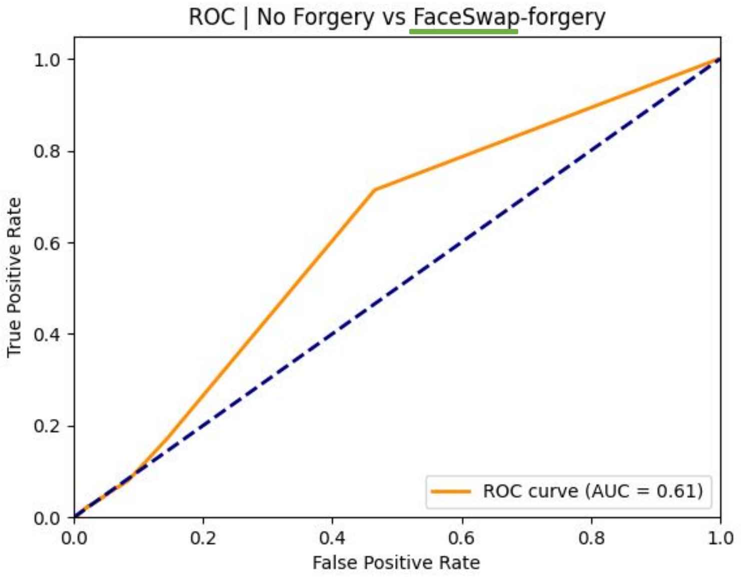
DeepFake Detection | Metric: AUROC → Real vs Fake Results

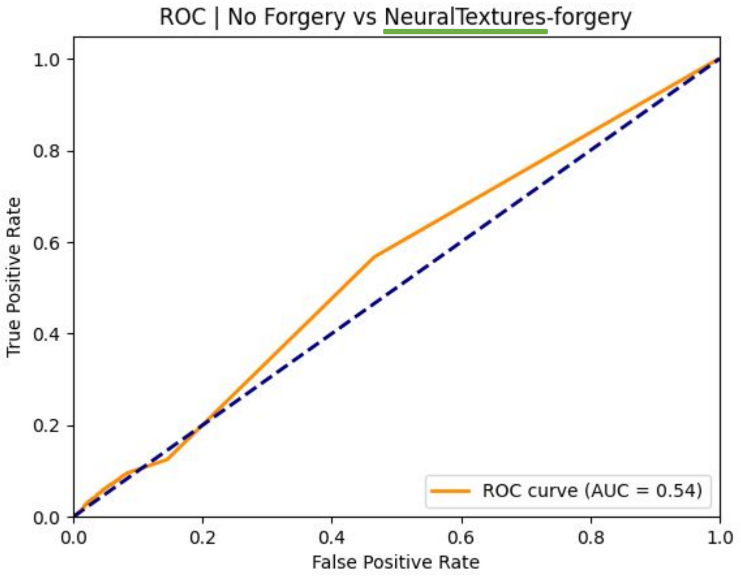
AUROC → AUC calculated on ROCs

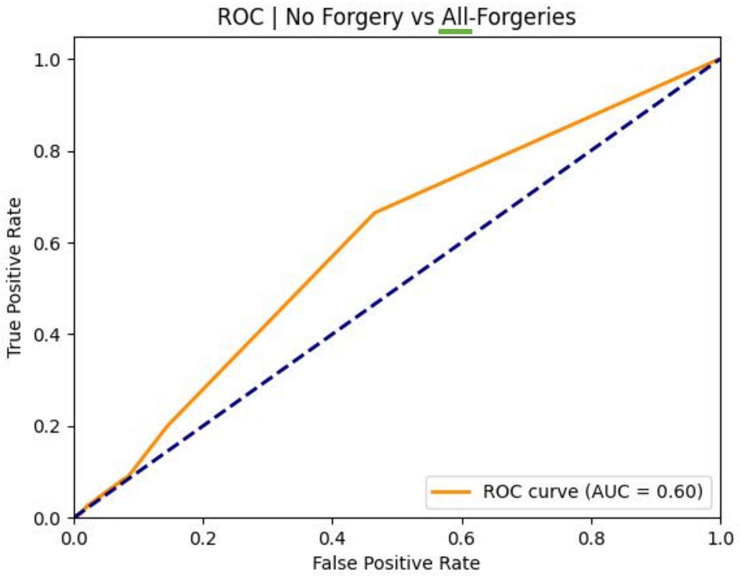

ROC | No Forgery vs Deepfakes-forgery


ROC | No Forgery vs Face2Face-forgery


ROC | No Forgery vs FaceShifter-forgery


ROC | No Forgery vs FaceSwap-forgery

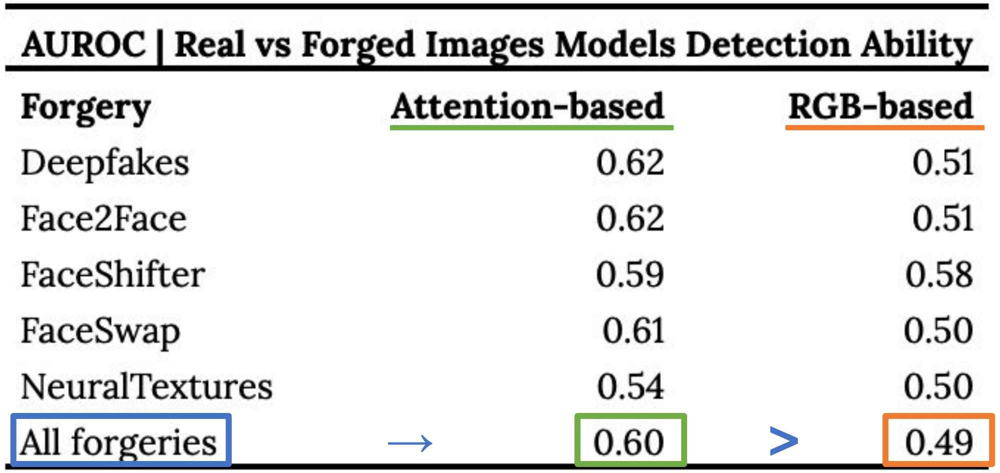

ROC | No Forgery vs NeuralTextures-forgery


ROC | No Forgery vs All-Forgeries

DeepFake Detection | Conclusions & Future Developments

On the method proposed

- Validated
- Independent from specific forgery
  - Transfer learning to new forgeries
- Performance
  - Better than RGB-based - random chance
  - Minor than Supervised SotA methods



AUROC | Real vs Forged Images Models Detection Ability

Forgery	Attention-based	RGB-based
Deepfakes	0.62	0.51
Face2Face	0.62	0.51
FaceShifter	0.59	0.58
FaceSwap	0.61	0.50
NeuralTextures	0.54	0.50
All forgeries	→ 0.60	> 0.49

Future advancements

- Try on ViT for demographic classification
  - *E.g. MiVOLO: Multi-input Transformer for Age and Gender Estimation*

Thanks for your Attention

REFERENCES: Images have been gathered from [See next slide]

REFERENCES:

1. Luca Cultrera, Lorenzo Seidenari, and Alberto Del Bimbo | "Leveraging Visual Attention for out-of-Distribution Detection" | Proceedings of the IEEE/CVF International Conference on Computer Vision | 2023
2. Zhong, Yaoyao, and Weihong Deng | "Face transformer for recognition" | arXiv preprint arXiv:2103.14803 | 2021
3. Kuprashevich, Maksim, and Irina Tolstykh | "MiVOLO: Multi-input Transformer for Age and Gender Estimation" | arXiv preprint arXiv:2307.04616 | 2023
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. | 2017 | “Attention is All You Need” | In Advances in Neural Information Processing Systems (pp. 5998-6008)
5. Alexey Dosovitskiy∗,† , Lucas Beyer∗ , Alexander Kolesnikov∗ , Dirk Weissenborn∗ , Xiaohua Zhai∗ , Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby∗,† ∗equal technical contribution, † equal advising Google Research, Brain Team | “An Image is Worth 16 x 16 Words: Transformers for Image Recognition at Scale” | at ICLR 2021
6. https://niessnerlab.org/projects/roessler2018faceforensics.html
7. https://medium.com/@birla.deepak26/autoencoders-76bb49ae6a8f
8. https://seeflection.com/22579/viral-deepfake-on-tiktok-causes-outrage/
9. https://towardsdatascience.com/using-transformers-for-computer-vision-6f764c5a078b
10. https://encord.com/blog/what-is-out-of-distribution-ood-detection/