Interpretable Spatial-Temporal Video Transformer for Deepfake Detection (ISTVT)

EE656A - Course Project Presentation

Durgesh Dongre (241040020)

Richik Majumder (241040068)

Saurabh Srivastava (231030609)

Anup Kumar (241010076)

Objective

- Implement and evaluate ISTVT: a transformer-based deepfake detector
- Explore interpretability of attention in video forensics

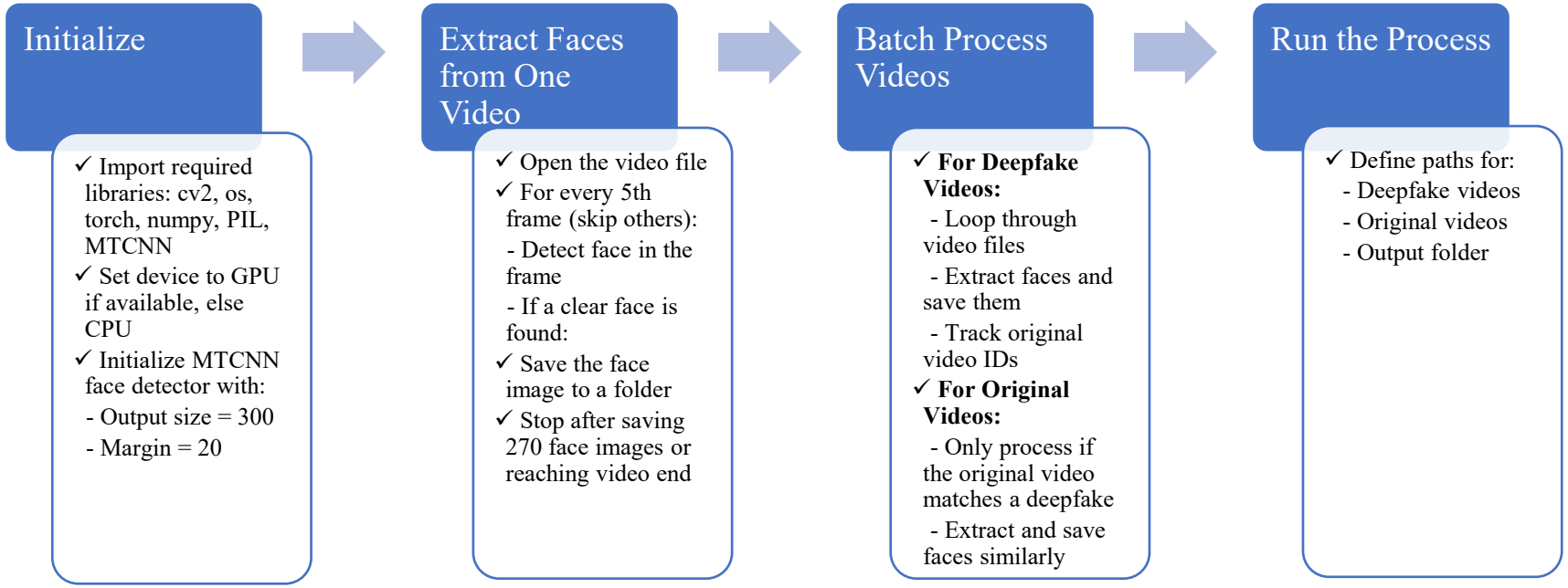
Dataset

- FaceForensics++ (Subset): 200 Deepfake videos
- Preprocessing: MTCNN face extraction, resized to 128x128
- The model was trained for 5 epochs using video data in C23 compression format, consistent with the FaceForensics++ dataset.
- The goal was to implement the full ISTVT architecture, but the temporal transform module was not included in this version of the model.

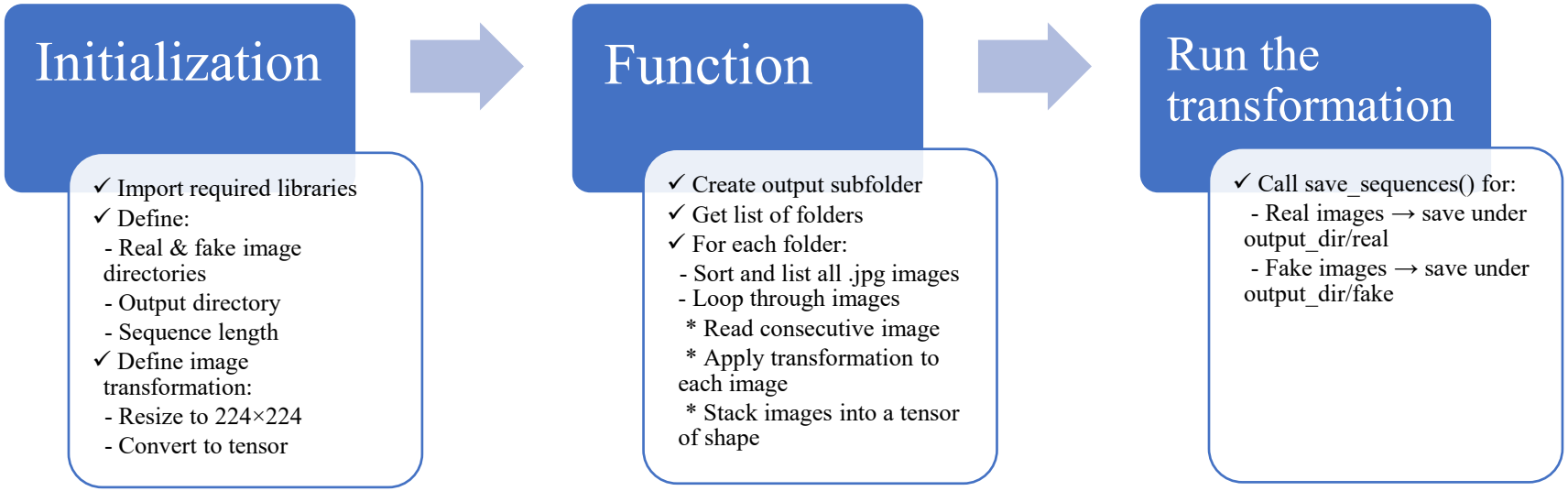
ISTVT Architecture Overview

- **Input**: Video frames → Xception CNN (feature extraction).
- **Tokenization**: Split features into patches → tokens.
- **Transformer Blocks**:
- **Spatial Self-Attention** (within-frame).
- **Temporal Self-Attention** (across-frame).
- **Classification Head:** Predicts "real" or "fake".

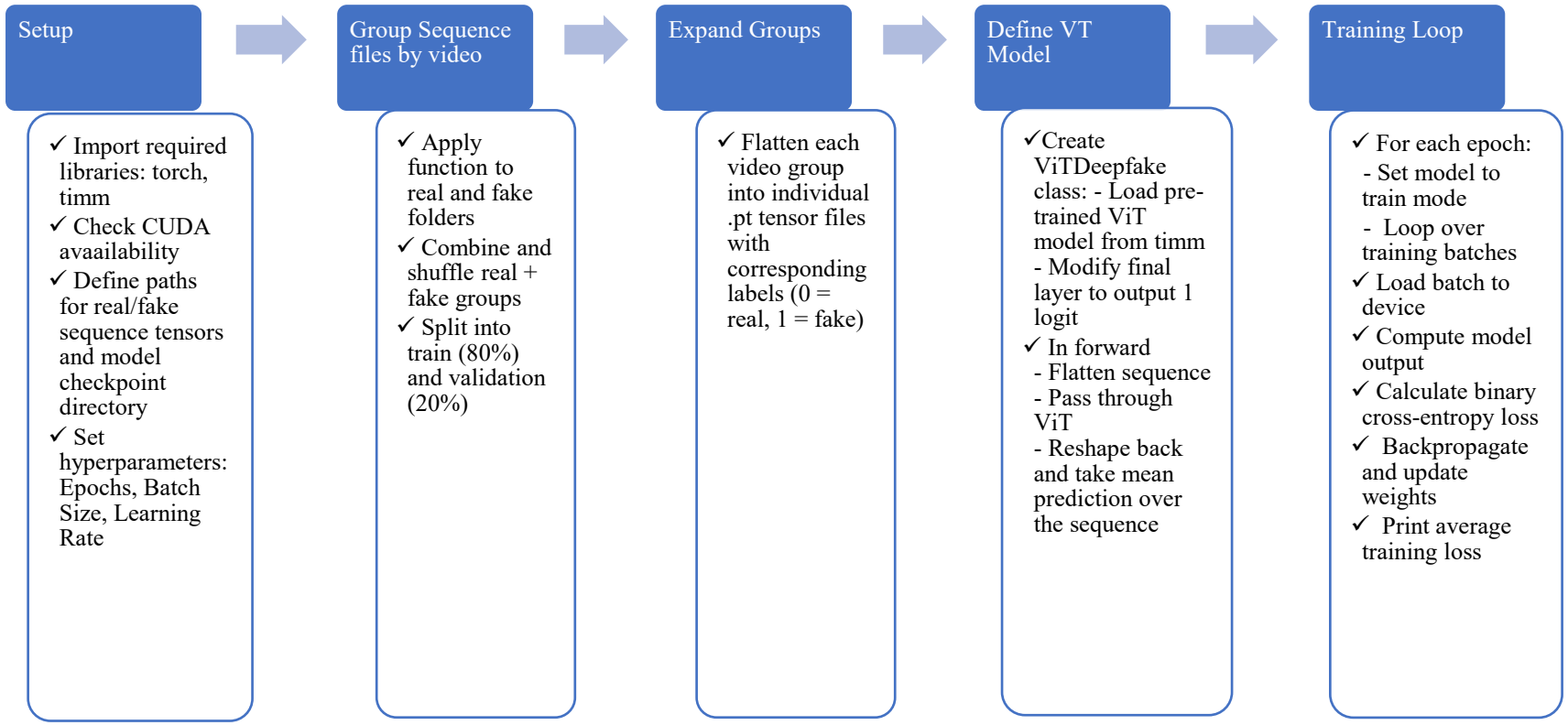
Pseudo-code for Face Extraction and Preprocessing



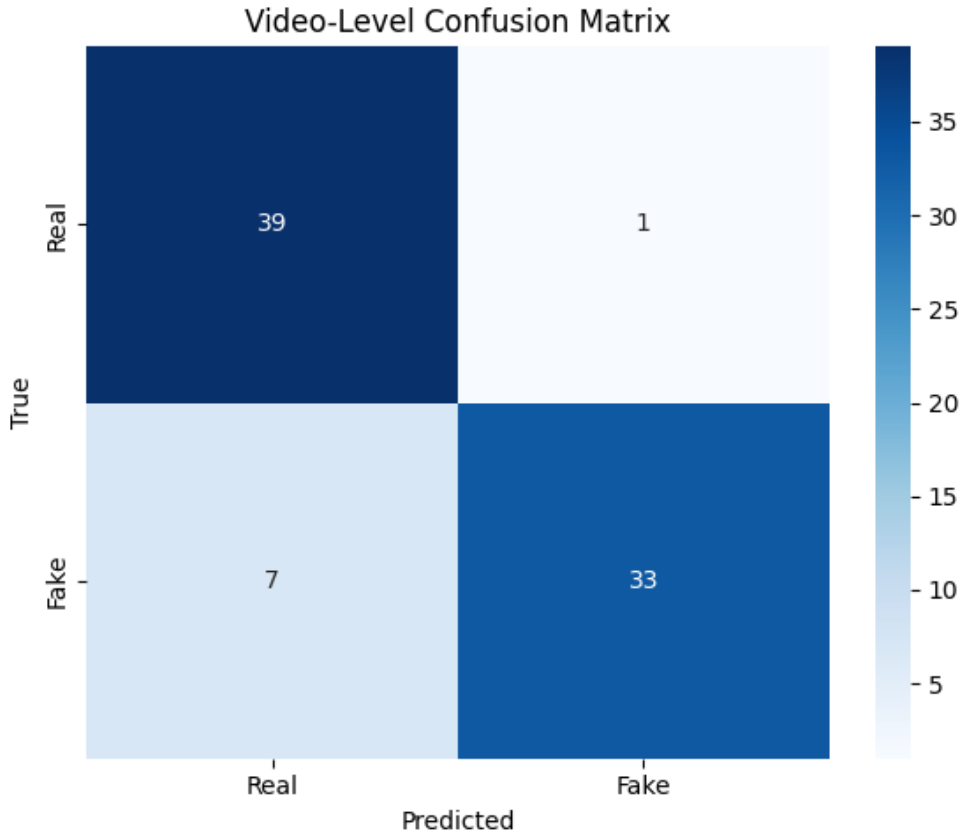
Pseudo-code for Image-to-Tensor Sequence Conversion



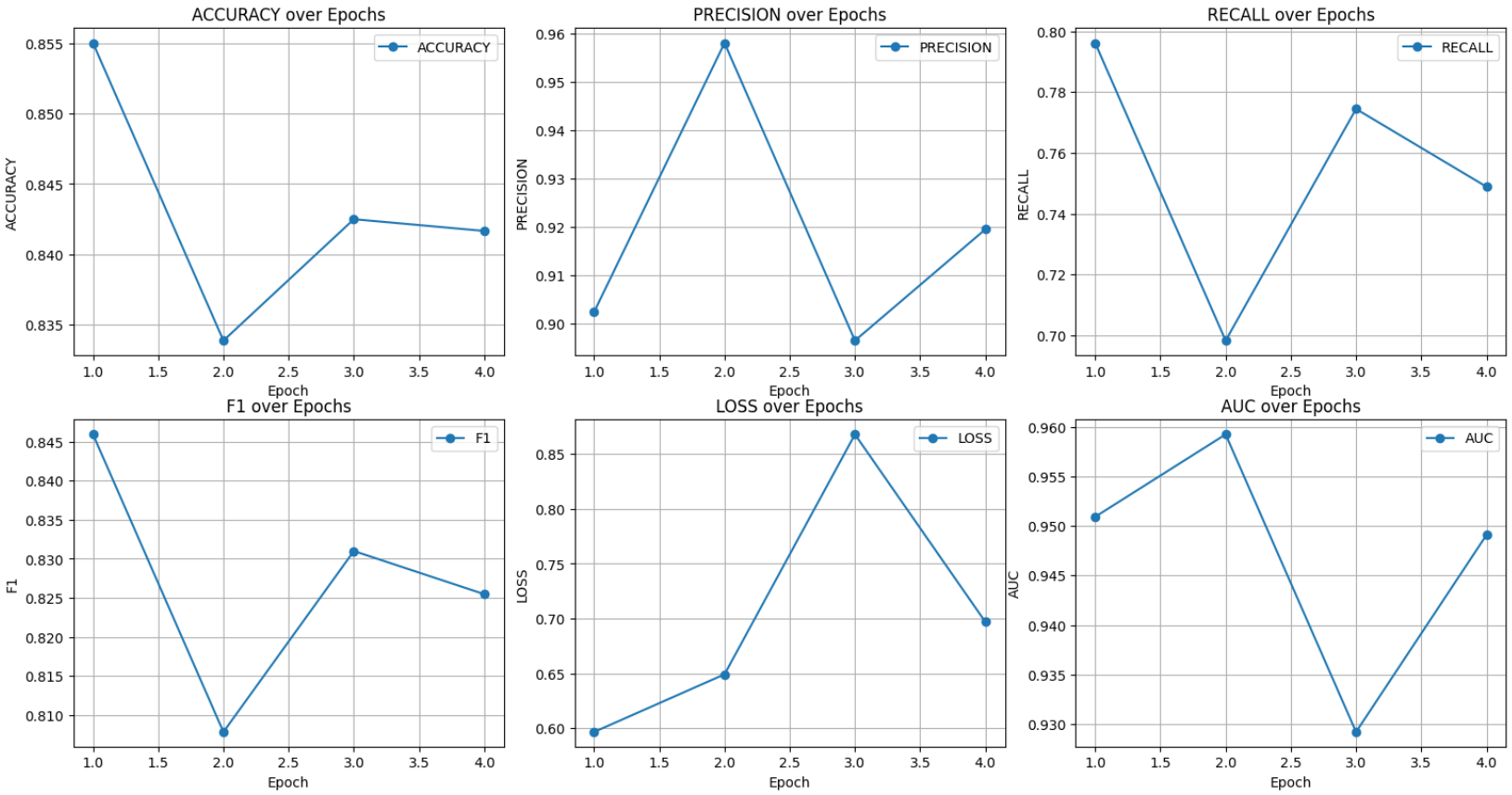
Pseudo-code for Vision Transformer for Deepfake Detection



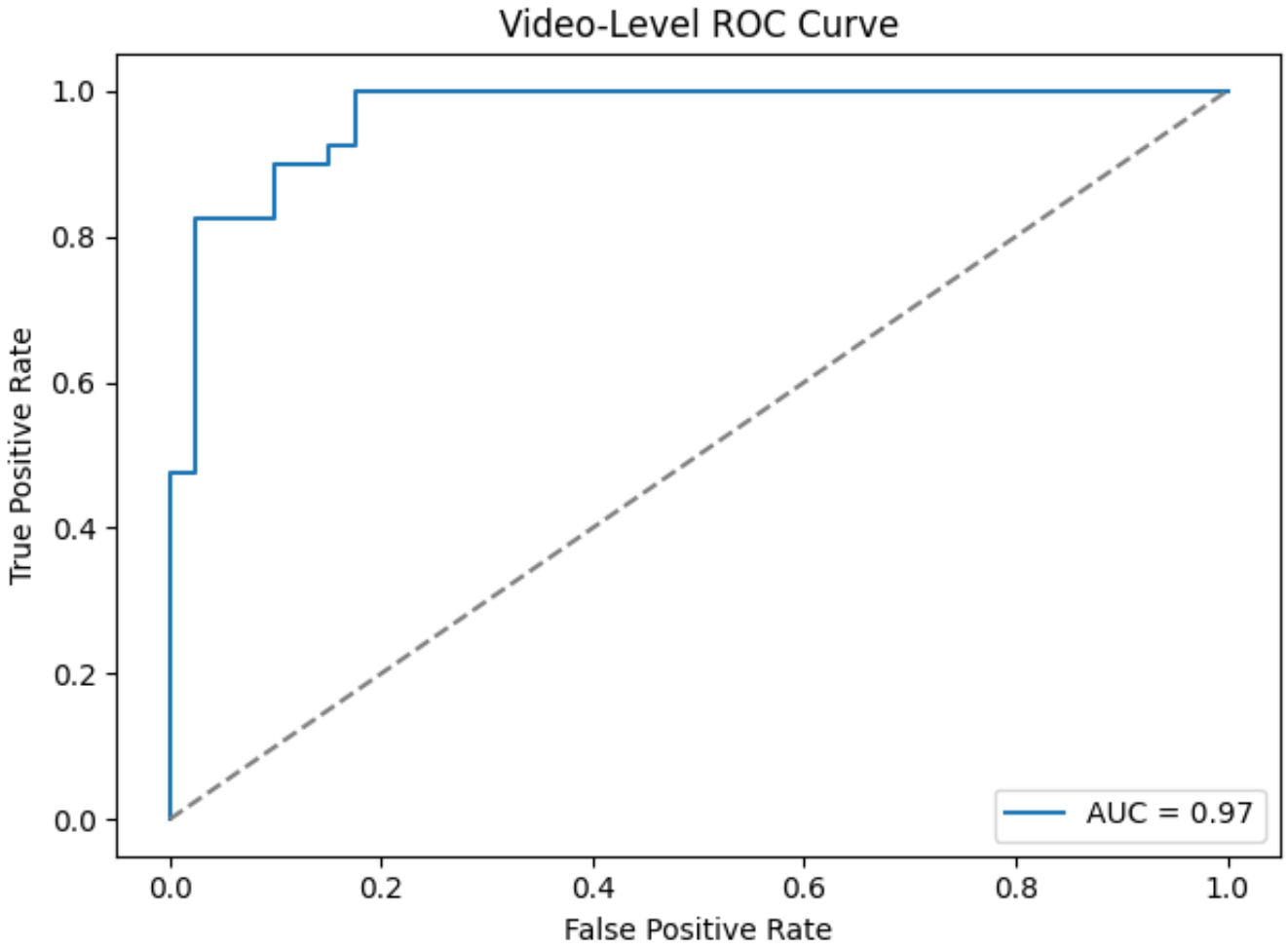
Results


Video-Level Confusion Matrix
Real
Fake
True
39
1
7
33
Real
Fake
Predicted
35
30
25
20
15
10
5

Evaluation Metrics Over Epochs



Video-Level ROC Curve



Conclusion

- The model effectively detects deepfakes with 90% accuracy and an AUC of 0.97, demonstrating robust performance.
- Future work could enhance model sensitivity to challenging samples.

Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.847826	0.975	0.906977	40
1	0.970588	0.825	0.891892	40
Accuracy	0.9	0.9	0.9	0.9
Macro Avg	0.909207	0.9	0.899434	80
Weighted avg	0.909207	0.9	0.899434	80

References

1. Zhou, X., Ding, Y., Liu, Y., Yu, N., & Liang, W. (2023). ISTVT: Interpretable spatial-temporal video transformer for deepfake detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 12345–12355.

2. Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1–7.

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

4. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? International Conference on Machine Learning (ICML), 139, 813–824.

5. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos. 2019 International Conference on Biometrics: Theory, Applications and Systems (BTAS), 1–8.

Thank You