EE656: ARTIFICIAL INTELLIGENCE MACHINE LEARNING DEEP LEARNING & ITS APPLICATIONS

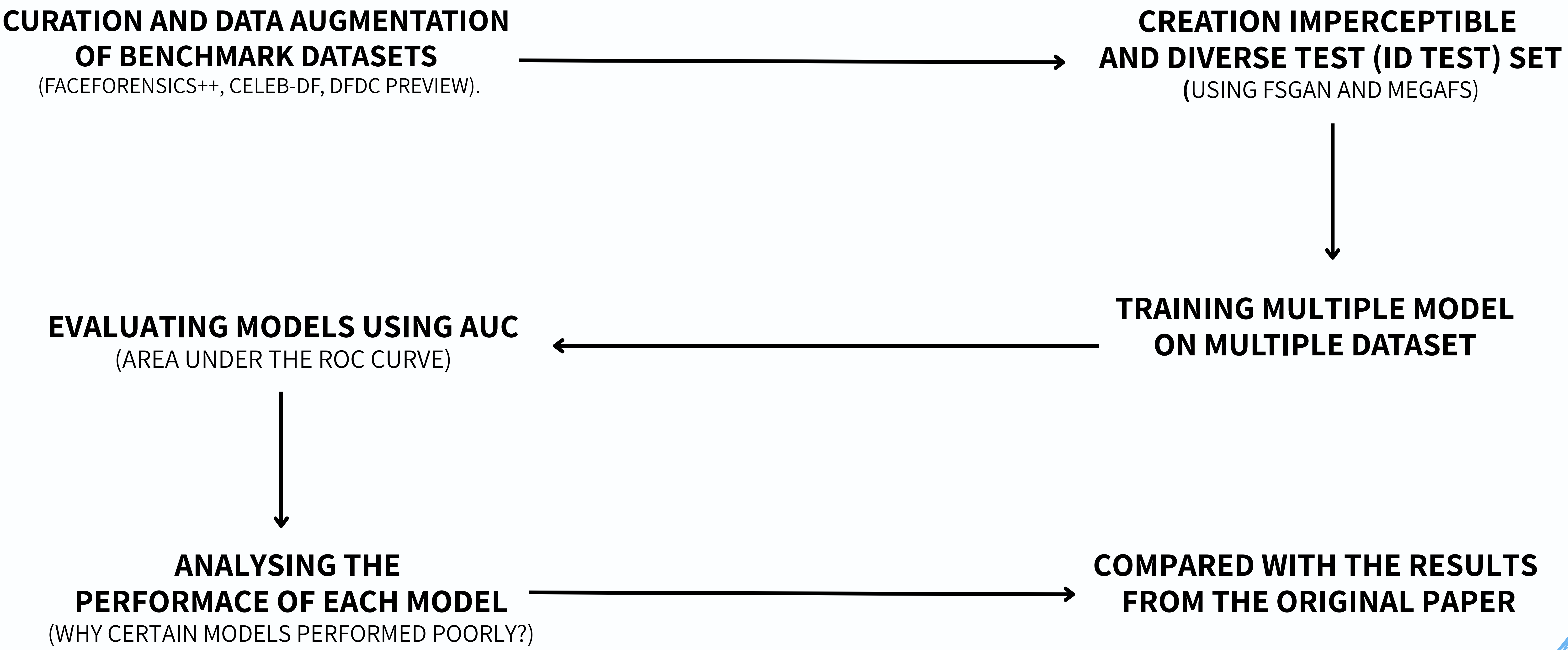
TOWARDS BENCHMARKING AND EVALUATING DEEPFAKE DETECTION

Group Members:

Amoghsiddhu R N	**(230127)**
Anish Kr. Mandal	**(230144)**
Vaishnavi Bhukya.	**(230295)**
Stuti Shukla.	**(231040)**

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

OVERVIEW



CURATION AND DATA AUGMENTATION OF BENCHMARK DATASETS
(FACEFORENSICS++, CELEB-DF, DFDC PREVIEW).

CREATION IMPERCEPTIBLE AND DIVERSE TEST (ID TEST) SET
(USING FSGAN AND MEGAFS)

TRAINING MULTIPLE MODEL ON MULTIPLE DATASET

EVALUATING MODELS USING AUC
(AREA UNDER THE ROC CURVE)

ANALYSING THE PERFORMACE OF EACH MODEL
(WHY CERTAIN MODELS PERFORMED POORLY?)

COMPARED WITH THE RESULTS FROM THE ORIGINAL PAPER

DATASET USED

1. UADFV

- A benchmark dataset with 49 real and 49 fake videos.
- Generated using FakeApp.
- Commonly used in early deepfake detection studies.

2. FaceForensics++ (FF++)

- Contains videos manipulated by multiple techniques: Deepfakes, Face2Face, FaceSwap, and NeuralTextures.
- Offers high-quality frames for training and evaluation.

3. DFDC (Deepfake Detection Challenge)

- Large-scale dataset by Meta AI.
- Includes thousands of real and fake videos.
- Offers diversity in actors, lighting, compression, and scenes.

**NOTE :- CELEB-DF, FORGERYNET, DF-TIMIT, AND DF-1.0 WERE EXCLUDED DUE TO THEIR LARGE SIZE AND HARDWARE/STORAGE LIMITATIONS.*

Imperceptible and Diverse Test (ID Test) Set

Purpose: Evaluate model performance on visually realistic and diverse deepfakes.
Sources: 7 public datasets + private dataset.
Private Dataset Generation:

- FSGAN (GAN-based): Official pretrained model.
- MegaFS (Autoencoder-based): Trained from scratch following original paper.
- Input sources: UADFV images (for FSGAN), FaceForensics++ raw videos (for MegaFS).

Selection Pipeline:

- Detection Filtering: Use trained Xception and Face-X-ray models to identify hard-to-detect (low score) fakes.
- User Perception Filtering:In the original paper, a group of 30 individuals was involved in selecting fake images that they mistakenly perceived as real to validate visual authenticity. However, due to constraints in time and resources, we could not replicate this large-scale human evaluation in our implementation.
- Final Result: ID Test Set containing imperceptible and diverse examples that challenge both data- and knowledge-driven forensic methods.

Imperceptible and Diverse Test (ID Test) Set

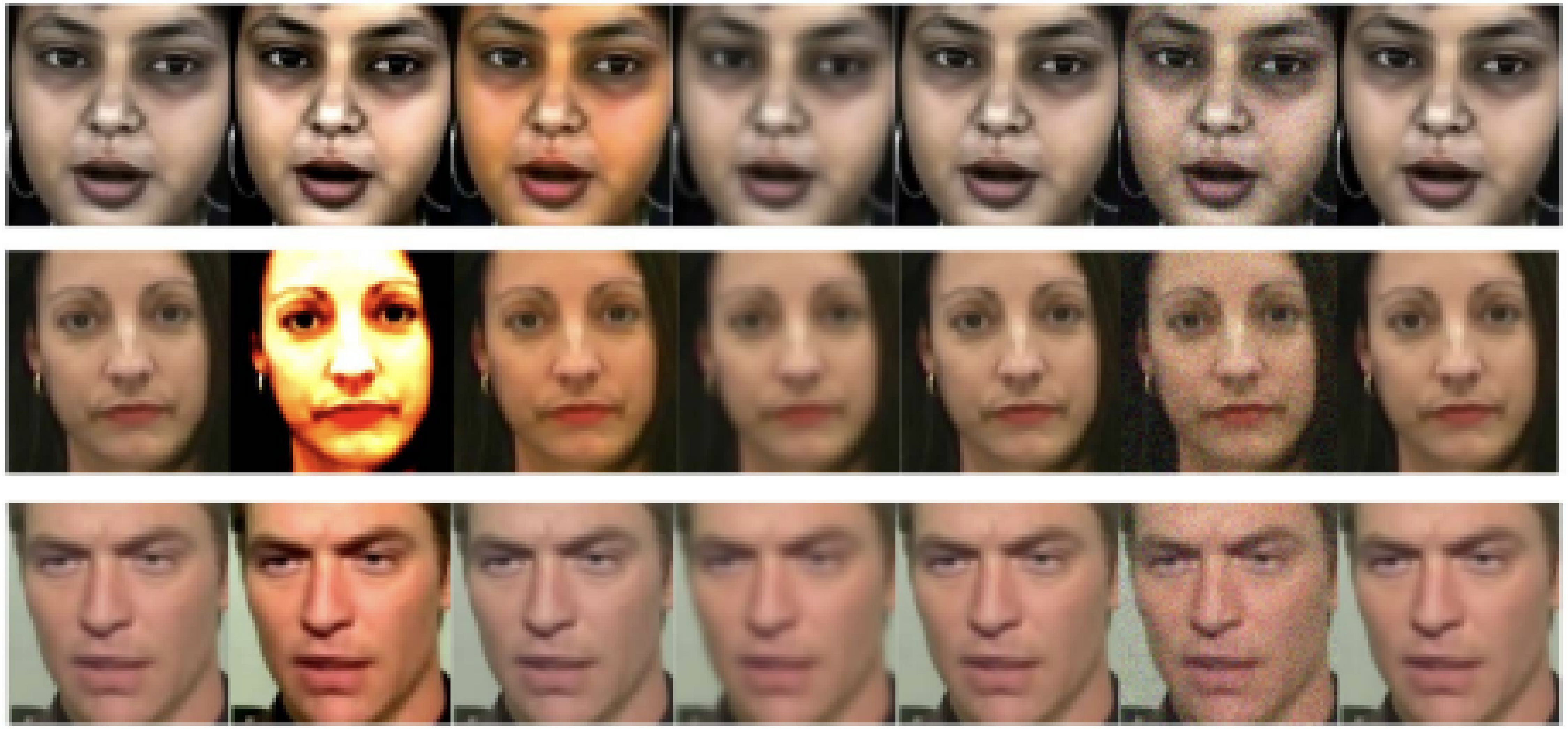
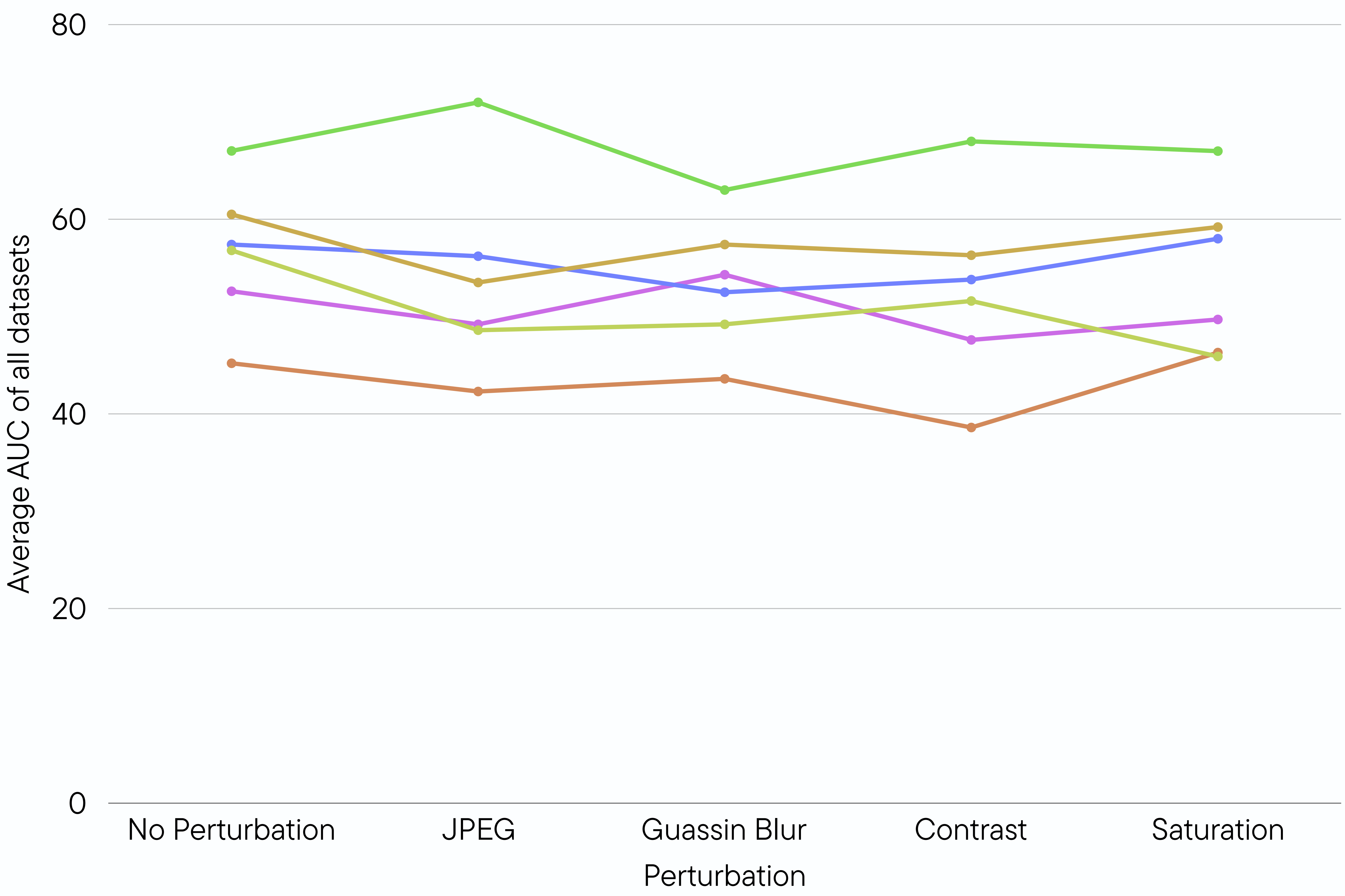


ILLUSTRATION OF FACE IMAGES IN ID TEST SET WITH DIFFERENT PERTURBATIONS. FROM LEFT TO RIGHT ARE RAW FACE IMAGES, FACE IMAGES WITH COLOR CONTRAST CHANGE PERTURBATION, COLOR SATURATION CHANGE PERTURBATION, GAUSSIAN BLUR PERTURBATION, JPEG COMPRESSION PERTURBATION, WHITE GAUSSIAN NOISE PERTURBATION, AND VIDEO COMPRESSION PERTURBATION.

Method	category	open source	Data Processing	Backbone	Params(M)	GFLOPs	Infer Time(ms)
FFD	Frame-Know	Yes	Detect + Add	XceptionNet + Reg. Map	20.82	16.84	6.04
Multiple-attention	Frame-Know	Part	Detect + Align	EfficientNet-b4	18.83	6.80	25.48
M2TR	Frame-Multi	Yes	Detect + Align	Eff-b4 + Transformer + Freq. Filter	18.61	2.92	34.18
Xception	Frame-Data	Part	Detect	XceptionNet	20.81	16.84	6.05
Mesonet-4	Frame-Data	Part	Detect + Align	4-layer Conv	0.28	0.12	5.11
HeadPose	Frame-Know	Yes	No	SVM	-	-	159.70
Patch Resnet Layer1	Frame-Data	Yes	Detect + Align	ResNet18	0.15	2.10	0.73

TRAIN TEST	UADFV	FF++/ DF	F++/ F2F	F++/ FS	F++/ FSwap	DFDC
HeadPose	80.99	50.01	48.67	54.06	47.89	76.02
Xception	89.20	95.45	96.35	93.25	96.57	76.48
FFD	88.35	93.56	92.49	89.34	95.23	74.38
Multi-Attention	99.93	95.31	71.65	88.45	85.93	-
M2TR	99.94	94.38	79.70	97.36	90.30	-
Mesonet-4	97.59	46.18	56.56	42.94	47.99	-
Patch Resnet Layer1	50.46	52.86	61.38	49.81	57.21	-


80
60
40
20
0
Average AUC of all datasets
No Perturbation
JPEG
Guassin Blur
Contrast
Saturation
Perturbation

Why models fail on the FF++ Dataset.

The FaceForensics++ (FF++) dataset is available in two formats: the original high-quality version (~1.2 TB) and a compressed version (~500 MB). In our evaluation, the compressed version was used, which has a significant impact on model performance due to the loss of critical low-level visual cues like textures, edges, and subtle artifacts.

This degradation particularly affects models such as MesoNet-4 and Patch ResNet Layer1, which rely on these fine-grained features for accurate deepfake detection. As a result, they tend to produce lower AUC scores on the compressed dataset—not necessarily due to weaker architecture, but because the compression masks manipulation traces, making detection more challenging.

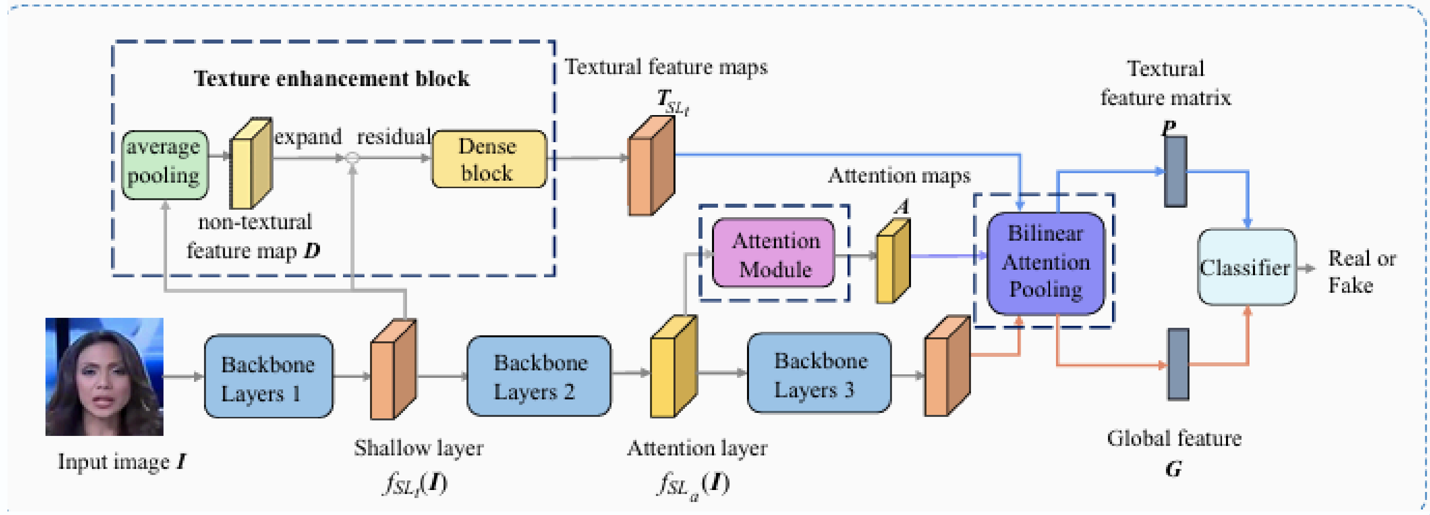
This introduces a bias in performance evaluation, especially for models focusing on local artifacts. For fair and meaningful comparisons, it's important to either use the original dataset or implement compression-aware training. Additionally, researchers should always report the dataset version used in their experiments to maintain transparency and ensure reproducibility.

Multi-Attention Deepfake Detection Model

- The Fine-Grained Multi-Attentional Deepfake Detection Network is a CNN-based model for frame-level binary classification of real and fake frames. It uses multiple attention mechanisms to focus on subtle local artifacts, treating deepfake detection as a fine-grained task. This attention-guided approach helps the model detect manipulation more effectively.

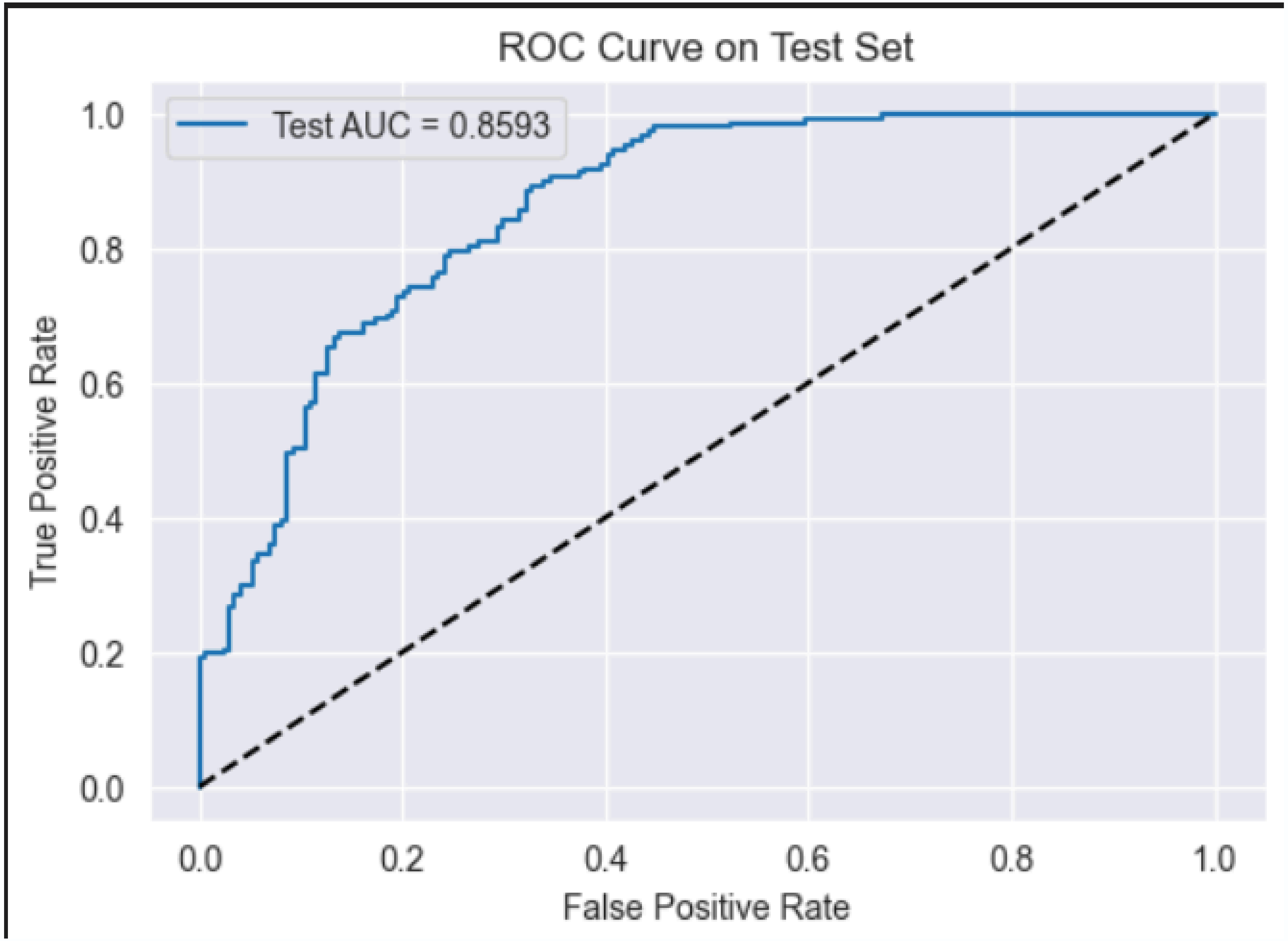
Model Architecture:

- The model uses EfficientNet-b4 to extract low-level (SLt) and high-level (SLa) features. It applies multiple attention modules to generate spatial maps that highlight subtle forgery cues, while dense convolutions refine shallow texture details to expose fine-grained artifacts.
- Bilinear Attention Pooling (BAP) fuses attention maps with multi-level features for classification. To enhance performance, the model uses Regional Independent Loss for diverse region focus and AGDA to boost generalization through attention-guided augmentation.



Training and Implementation:

- The model processes a single normalized RGB frame (380×380) through EfficientNet-b4, using Adam/SGD (LR starting at 1e-4) and batch sizes of 16–32. It trains with a combined loss of binary cross-entropy, regional independent loss, and AGDA, passing features through attention modules and BAP to output a real/fake score, with attention maps for interpretability.



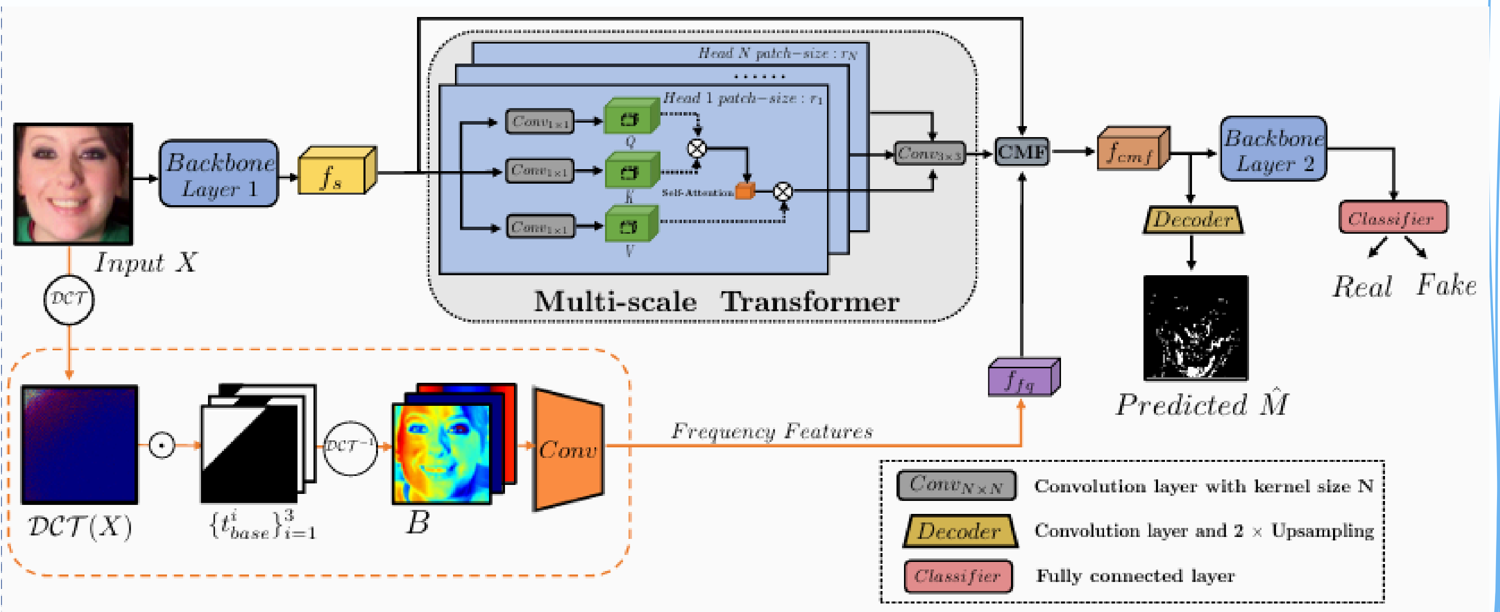
TEST AUC: 0.8593

M2TR: Multi-stream Multi-scale Transformer

- M2TR is categorized as an intra-frame level, multi-stream-driven detection model, designed specifically to capture subtle and diverse patterns associated with facial forgeries.

Model Architecture:

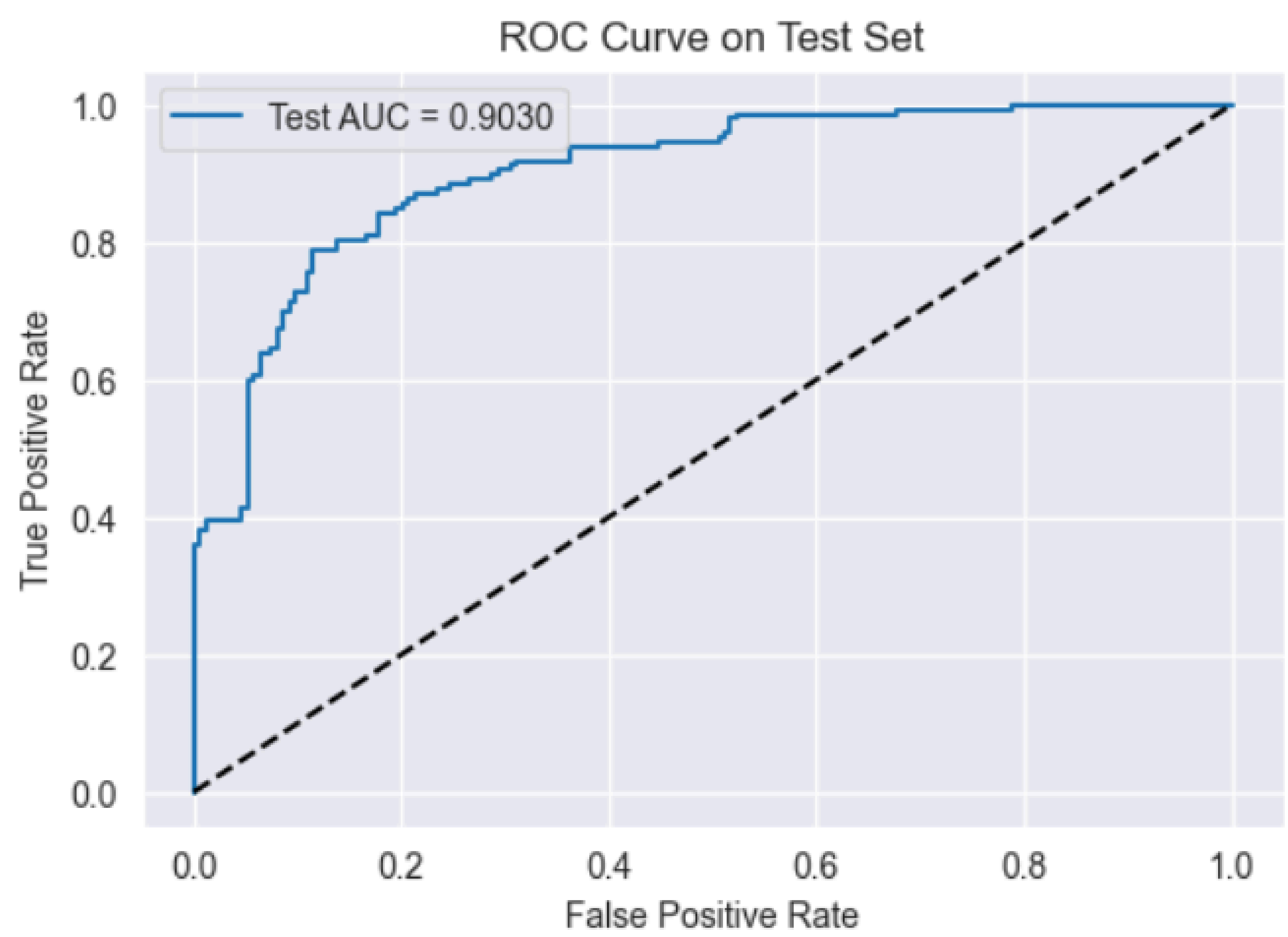
- M2TR uses a dual-stream architecture, where the spatial stream leverages multi-scale transformers to capture detailed spatial forgery patterns, and the frequency stream uses frequency filters to detect subtle manipulation traces.
- These complementary features are fused through a cross-modality fusion block to enhance the model's deepfake detection accuracy.



Overview of the proposed M2TR. The input is a suspicious face image (H x W x C), and the output includes both a forgery detection result and a predicted mask (H x W x 1), which locates the forgery regions.

Training and Implementation:

- M2TR is trained using only classification loss, in accordance with its official implementation, without any auxiliary loss functions.
- The model is trained with hyperparameters and configurations consistent with the original paper to ensure reproducibility, and all images are uniformly preprocessed using Dlib for face detection, alignment, and cropping to maintain experimental fairness across all evaluated models.



Test AUC: 0.9030

Xception Deepfake Detection Model

- Xception is categorized as an intra-frame level data driven method. This method adopts XceptionNet as the backbone for binary classification.

Model Architecture:

- Xception is an intra-frame level, data-driven deepfake detection model that uses XceptionNet as its backbone—a deep convolutional network built on depthwise separable convolutions for efficient and powerful feature extraction.
- The model is designed to extract fine-grained spatial features from individual video frames and acts as a binary classifier, predicting the probability of an image being real or fake based on facial manipulation cues.

Training and Implementation:

- The Xception model is trained using a batch size of 32 with the Adam optimizer and a learning rate of 0.0002, optimizing a binary classification loss to distinguish real from fake images, in line with the original implementation.
- All input images are uniformly preprocessed using Dlib for face detection, alignment, and cropping, ensuring consistency and fairness across all models evaluated in the benchmark.

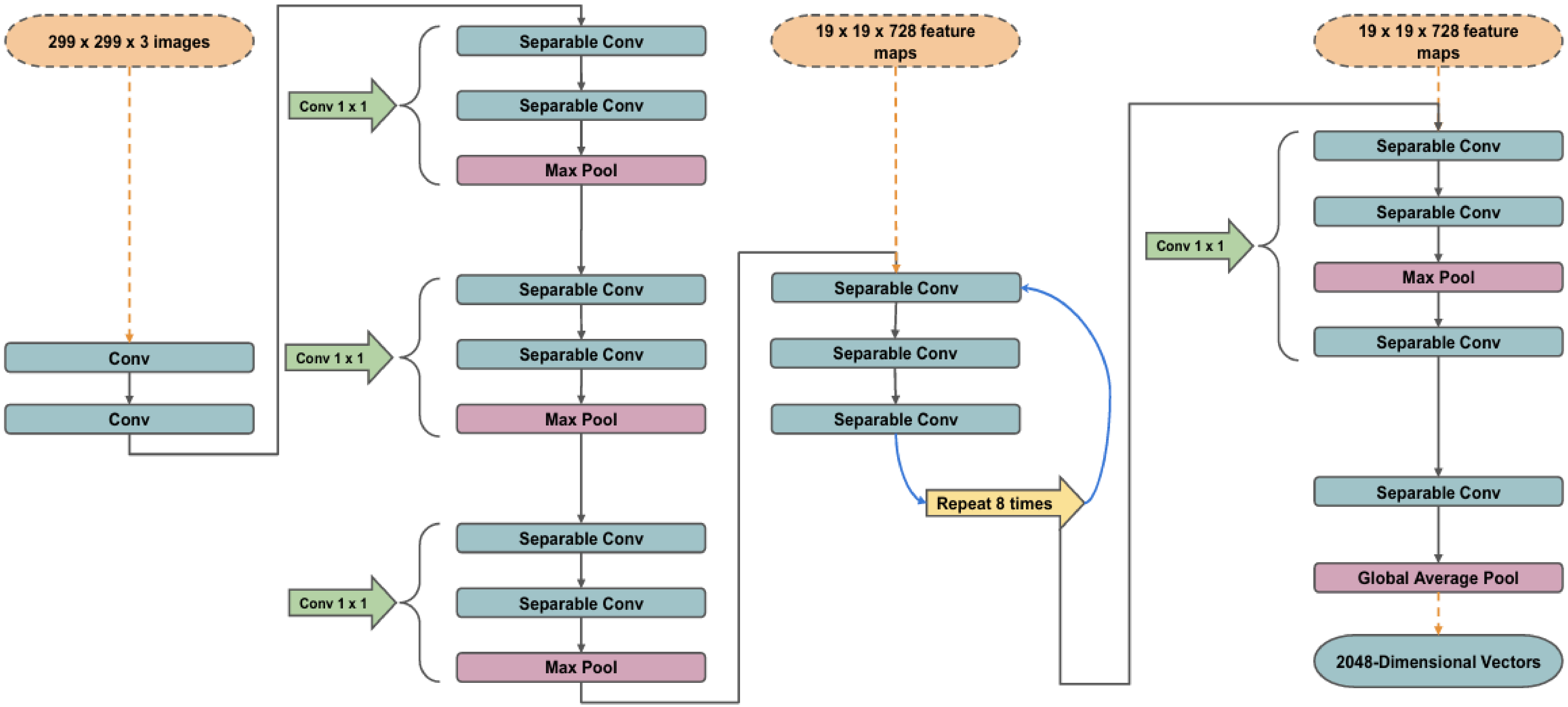


Figure 3. Architecture of Xception

MesoNet-4

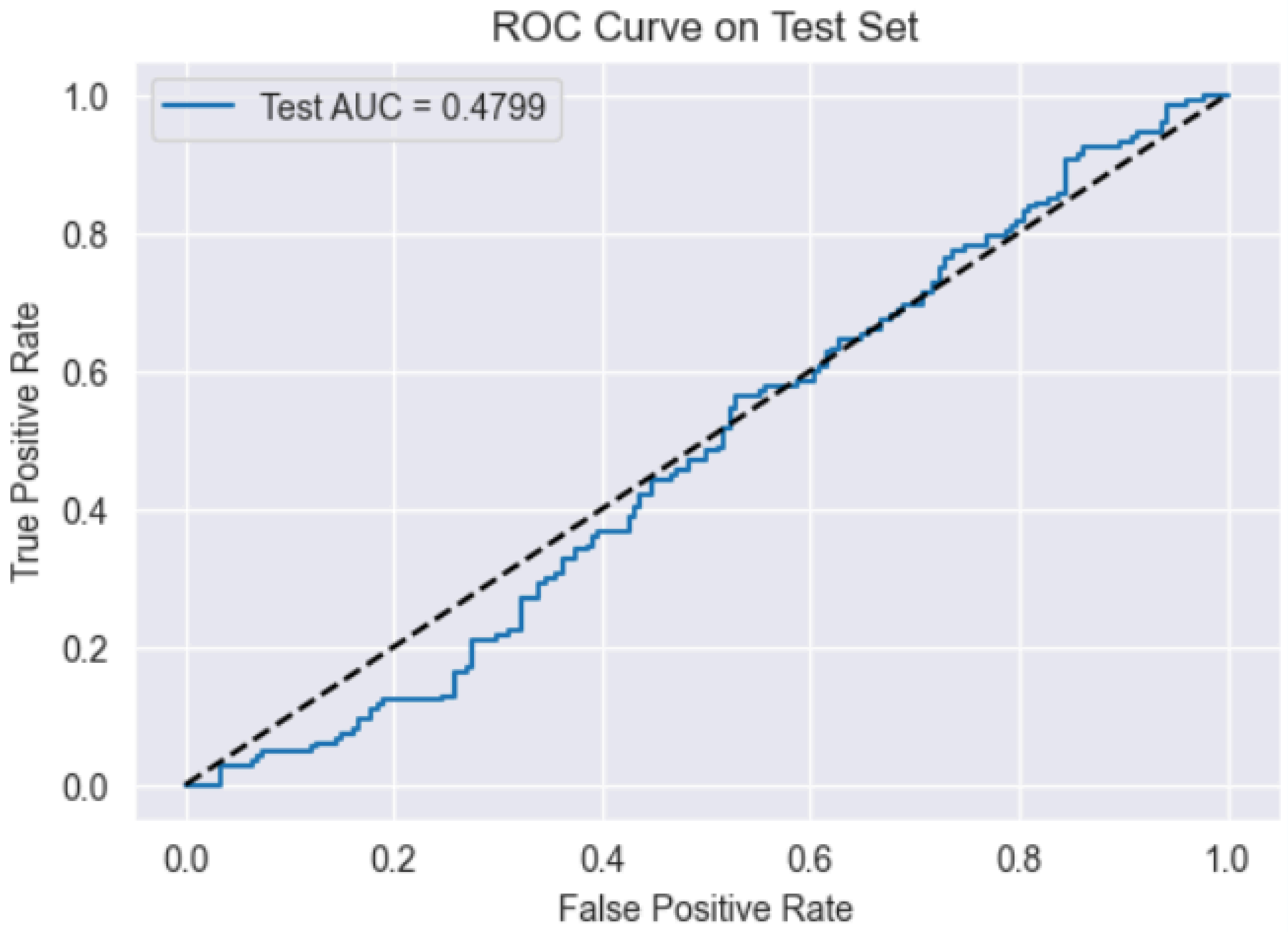
- MesoNet-4 and MesoInception-4 are shallow CNN models used for frame-level binary classification of real and fake video frames. They follow a data-driven approach, focusing on mesoscopic (mid-level) features rather than raw pixels or high-level semantics. These models operate at the intra-frame level, processing each frame independently to detect manipulation with low computational cost.

Model Architecture:

- MesoNet-4 is a shallow CNN with four conv-pool layers and a dense layer, designed for speed and focused on mid-level features often missed by deeper models.
- MesoInception-4 enhances MesoNet-4 by replacing the first two layers with an Inception module using dilated convolutions to capture multi-scale features.

Training and Implementation:

- MesoNet-4 processes resized and normalized RGB frames using the Adam optimizer with a learning rate decaying from 1e-3 to 1e-6, and a batch size of 75. It is trained with binary cross-entropy loss to perform frame-level real/fake classification, outputting a probability score per frame, which can be averaged across frames for reliable video-level deepfake detection.



Test AUC: 0.4799

Patch ResNet Layer1

- The model performs frame-level binary classification using a patch-wise supervision approach, where each image is divided into patches and analyzed individually.
- It follows a data-driven intra-frame CNN strategy using truncated ResNet or Xception, focusing on learning local features through patch-level predictions and aggregating them for image-level decisions.

Model Architecture:

- Each input image is divided into multiple patches, and the model makes individual predictions for each patch based on learned local features.
- During testing, the patch-level outputs are averaged to generate the final image-level prediction, allowing the model to capture both local inconsistencies and overall manipulation patterns.

Training and Implementation:

- During training, images are split into patches and passed through truncated ResNet/Xception models, with patch-level labels guiding the network to detect local manipulations. The model uses patch-level loss (e.g., binary cross-entropy), a batch size of 32 (16 real, 16 fake), and the Adam optimizer. In testing, predictions are made per patch and then averaged to produce the final image-level real/fake score, with losses and metrics computed accordingly.

HeadPose Deepfake Detection Model

- HeadPose is an intra-frame, knowledge-driven method that detects deepfakes by analyzing 3D head pose differences between the full and central face regions, using an SVM classifier for final classification.

Model Architecture:

- HeadPose is an intra-frame, knowledge-driven deepfake detection method that relies on 3D head pose estimation using facial landmarks, offering an interpretable feature-based approach to forgery detection.
- The model extracts discriminative features by comparing 3D head pose differences between the entire face region and the central face region, which tend to show unnatural inconsistencies in deepfake images.

Training and Implementation:

- The extracted pose difference features are used to train a Support Vector Machine (SVM) classifier, which learns to distinguish between real and manipulated images based on these inconsistencies.
- As a lightweight, non-deep learning approach, HeadPose can be trained on smaller datasets, and it does not require GPU-based training pipelines, making it fast and efficient.

FFD (Face Forensics Detection)

- FFD is an intra-frame level CNN designed for frame-level binary classification and manipulation localization. It follows a knowledge-driven, attention-guided approach by integrating attention layers into the architecture. The model focuses on both detecting and localizing manipulated regions within individual video frames.

Model Architecture:

- FFD uses XceptionNet as its backbone, a deep CNN with depthwise separable convolutions, effective for face forensics and image classification tasks. It serves as the main feature extractor for detecting manipulation patterns.
- The model integrates an attention mechanism with two proposed modules: the Manipulation Appearance Model and Direct Regression. In this setup, Direct Regression is used to generate attention maps that guide the network to focus on discriminative regions, enabling both detection and fine-grained localization of manipulated areas.

Training and Implementation:

- FFD takes resized and normalized RGB frames as input and uses supervised attention learning to guide attention map training. It is optimized using Adam, with a combination of three losses: binary classification, attention map regression, and an auxiliary loss for stability. The model outputs a real/fake probability per frame along with attention maps that localize manipulated regions for interpretation.

THANK YOU