Detection of Deep-Morphed Deepfake Images to Make Robust Automatic Facial Recognition Systems

Presenter: Alakananda Mitra

A. Mitra[1], S. P. Mohanty[2], P. Corcoran[3], and E. Kougianos[4]

University of North Texas, Denton, TX , USA.[1,2,4] and

National University of Ireland, Galway, Ireland[3].

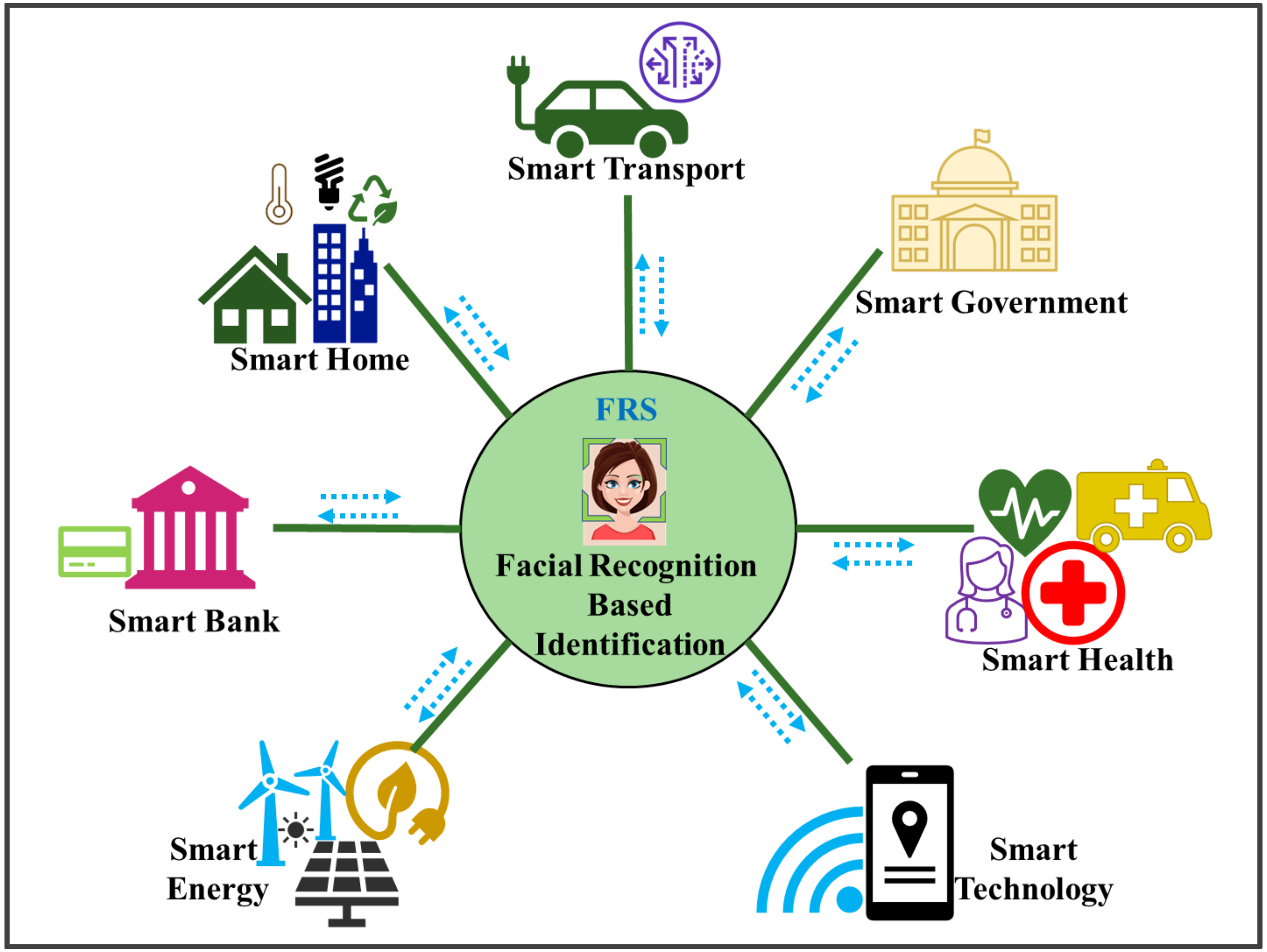
Email: alakanandamitra@my.unt.edu[1], saraju.mohanty@unt.edu[2], peter.corcoran@nuigalway.ie[3], and elias.kougianos@unt.edu[4]

Outline

- Facial Recognition System
- Attacks on Facial Recognition System
- Deep-Morphed Deepfake Attack
- Proposed Solution
- Results
- Conclusions & Future Work

image: Freepik.com

Identification of Individual in Smart City



Facial Recognition System (FRS)

- Facial Recognition System
  - Biometric based Identification System - Unique to the User
- Non-invasive Identification System - No Touching required
- Process of Identifying or Verifying the Identity of a Person using his/her Face
- Steps for FRS:
  - **Face Detection:** Detecting and Locating Human Faces in Images/ Videos
  - **Face Capture:** Changes Information (Features) of a face into a Set of Vectors
  - **Face Match:** Verifies if Two Faces are of the Same Person

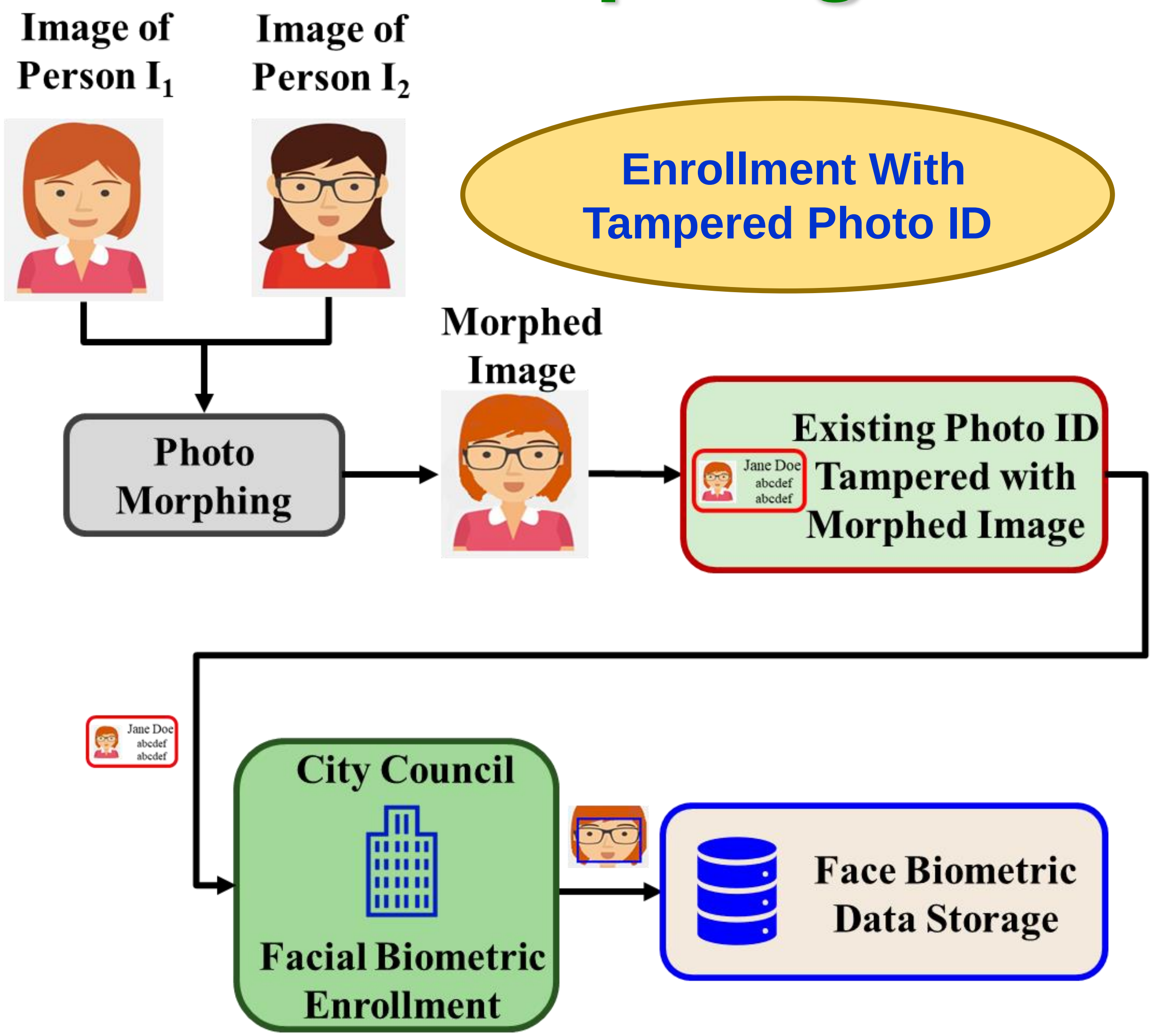
Attacks on FRS

- **Susceptible to Attacks**
  - **Presentation Attack : A Biometric Spoof Detected when Presented to** a Biometric Sensor
  - **Indirect/Channel Attack :** When Data Moves in the Network without Encryption
  - **Face Morphing Attack (FMA) :** Morphed Image
    - Traditional – Landmark Points Based
    - Deep-Morphed Deepfake – GAN Generated
      - (MorGAN, StyleGAN, FSGAN)

Deep-Morphed Deepfake

- **Deepfake = Deep Learning + Fake**
- Created by Deep Learning Networks
  - **Generative Adversarial Networks (GANs)**
- Sophisticated Images
- Make Face Morphing Easy and Realistic
- Rampant in Social Media and Websites
- Change the Perception of TRUTH
- Threat to Biometrics Based Facial Recognition Systems

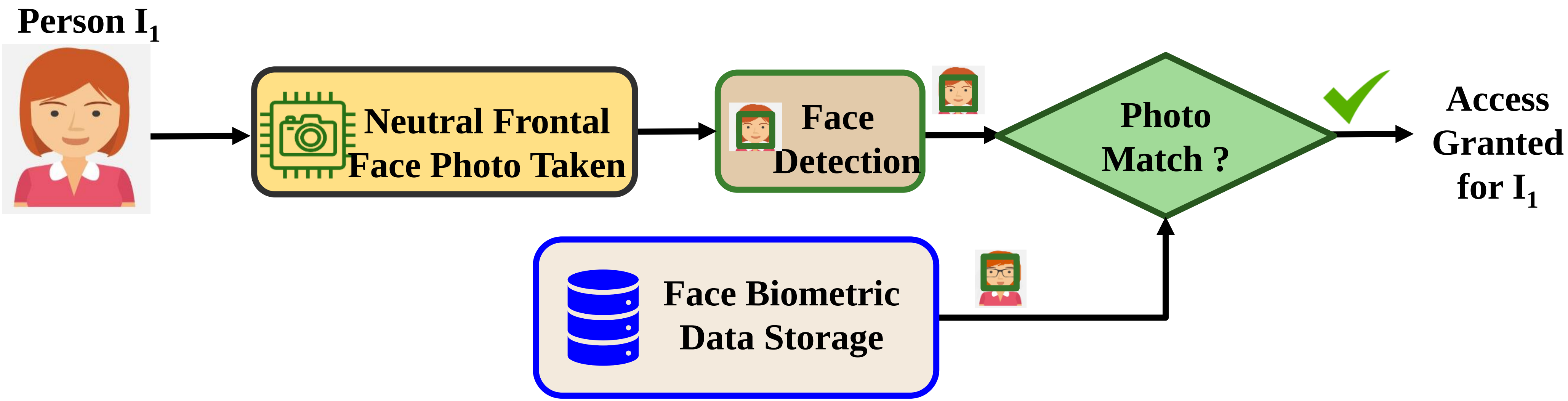
Face Morphing Attack on FRS of Smart City



- Citizens Submit their Existing Photo ID to the City Council Office
- Hostile – Person I1; Victim – Person I2
- ID of Hostile Person I1 Tampered with Morphed Photo from Victim Person I2
- Photo of the ID Matched with the Hostile Person I1
- Registered in the FRS

Photo ID Submission → **Facial Matching with Photo ID** → **Registering Biometric Facial Data from ID**

Face Recognition At Smart City Facility



- Hostile Person I1 comes to a Smart City Facility
- I1’s Face matches with the Data Stored
- Gains Access to that Facility

Proposed Solution For The Problem

Problems

- Misuse of FRS
- Innocent People Victims
- Hostile People take Advantages

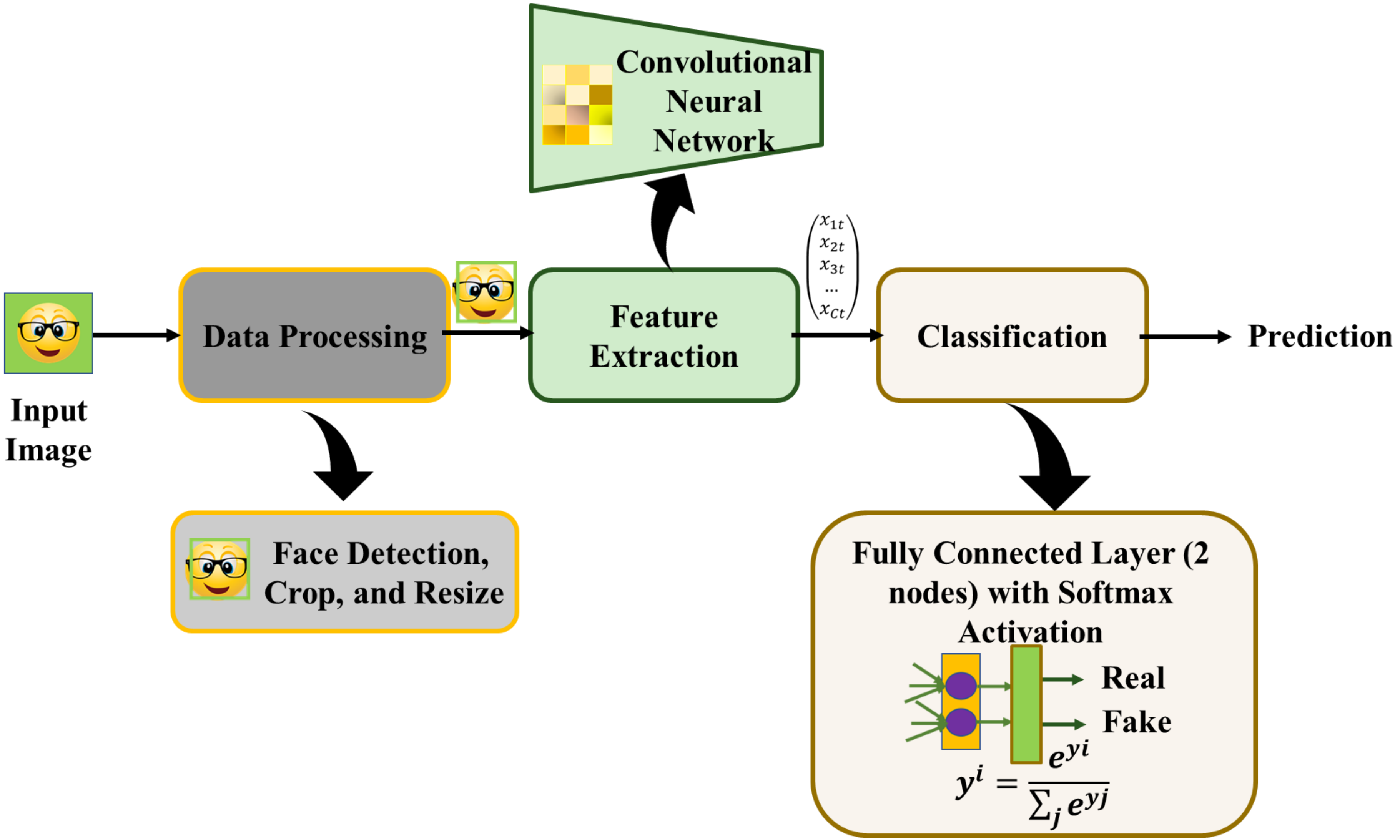
Solution Proposed

- CNN based network detects Deep-Morphed Deepfake Images
- Is used to detect images submitted for registration in FRS of smart cities
- Light Weight - IoT friendly model, makes the Registration Process easy and not localized to Council Office

Related Works

Papers	Dataset	Methods	AUC/ACC
Matern et al. [2019]	DeepfakeTIMIT	Visual Aspects + Logistic Regression + MLP	Low
Yang et al. [2019]	DeepfakeTIMIT	Head pose + Facial expression + dlib + SVM	Low
Afchar et al. [2018]	DeepfakeTIMIT	Mesoscopic Features	High
Zhou et al. [2018]	DeepfakeTIMIT	Steganalysis + Deep learning feature	Low
Nguyen et al. [2019]	DeepfakeTIMIT	Capsule network	Low
Proposed Method [2021]	DeepfakeTIMIT	**CNN based**	**Highest**

CNN Based Detection Method



CNN Based Detection Method (Contd..)

- MobileNet V2 as Feature Extractor
- Depthwise Separable Convolution
- Linear Bottleneck between layers
- Shortcuts connect the bottleneck layers
- Last FC layer with ImageNet classes changed to a FC layer with softmax activation and two nodes

CNN Based Detection Method (Contd..)

- MobileNet V2 as Feature Extractor
- Depthwise Separable Convolution
- Linear Bottleneck between layers
- Shortcuts connect the bottleneck layers
- Last FC layer with ImageNet classes changed to a FC layer with softmax activation and two nodes

Datasets

Real		Fake	
Data source	**# of Image**	**Data source**	**# of Image**
VidTIMIT	34,004	DeepfakeTIMIT (HQ)	33,988
VidTIMIT	34,004	DeepfakeTIMIT(LQ)	34,025

Data	# of Images		
	Real	**Fake**	
		DeepfakeTIMIT (HQ)	**DeepfakeTIMIT(LQ)**
Train	23,873	23,939	23,965
Validation	6,135	6,000	6,010
Test	3,996	4,049	4,050

DeepfakeTIMIT

- 32 subjects
- Total of 620 videos
- A lower quality (LQ) with 64x64 in/out size
- A higher quality (HQ) 128x128 in/out size
- Fake image frame rate 25 fps

VidTIMIT

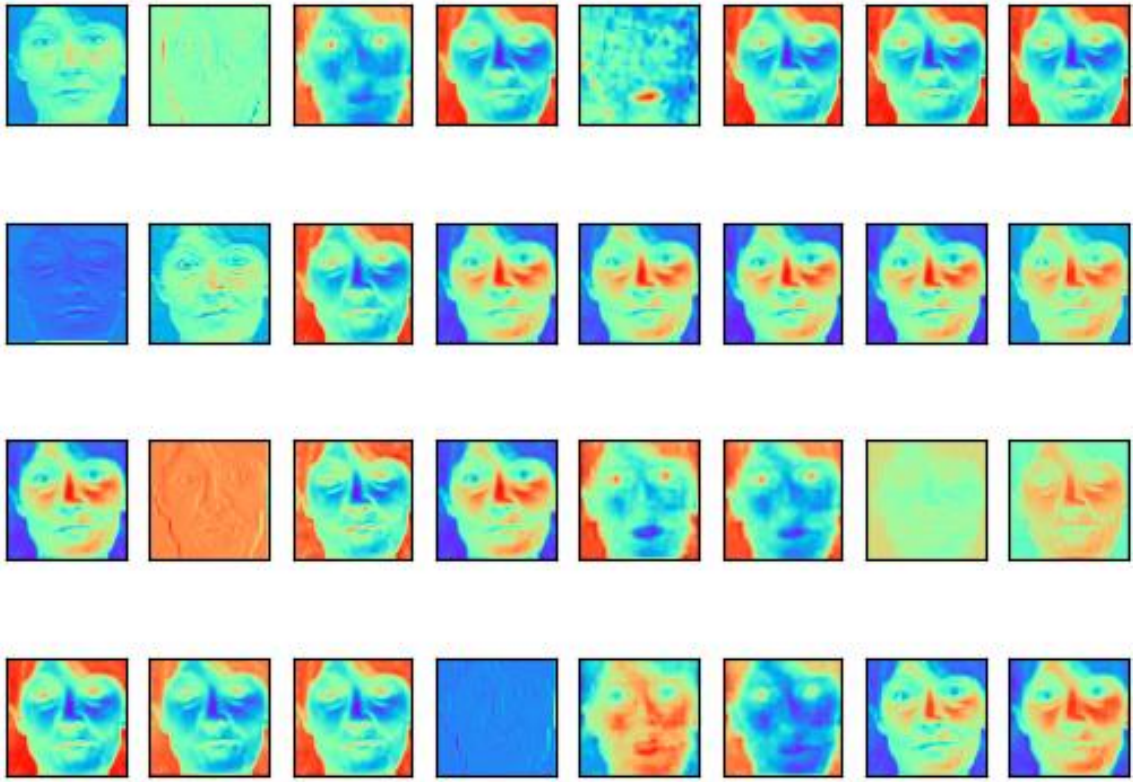
- Same subjects' videos as DeepfakeTIMIT

Implementation & Training Protocols

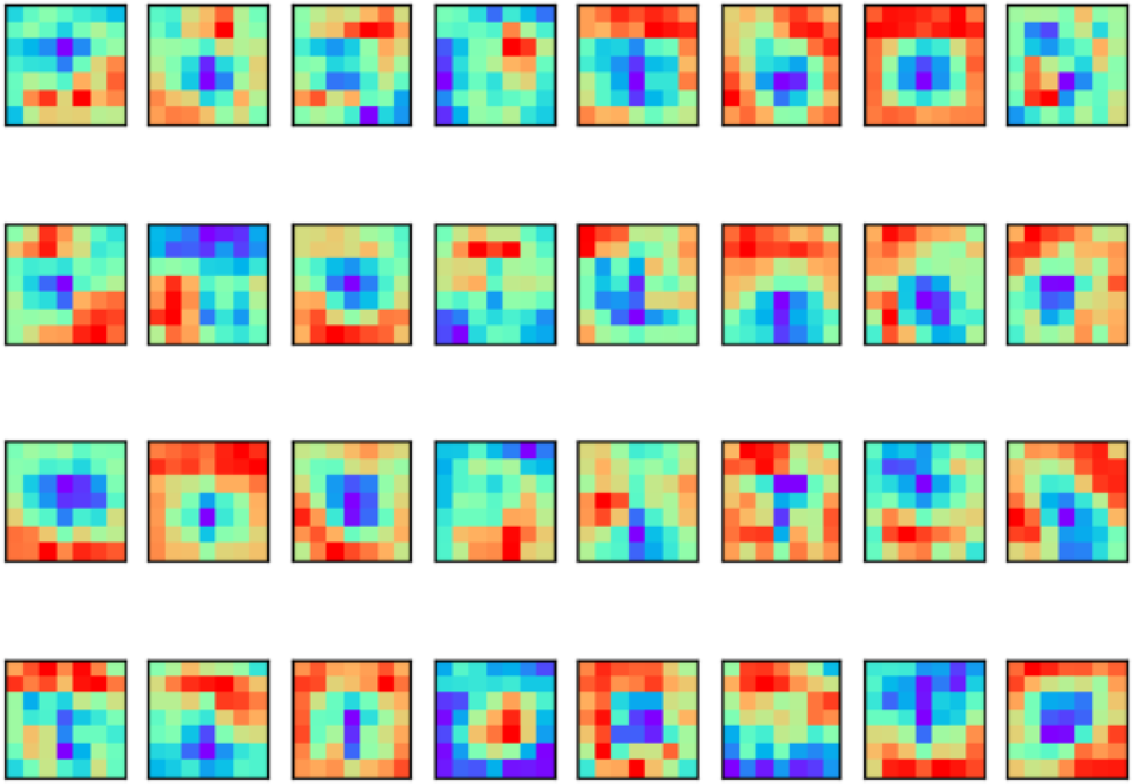
- Data Augmentation
- Transfer Learning
  - Accuracy Higher
  - Training Time Lower
- Feature Extractor kept frozen. Last FC layer trained for 10 epochs
- End-to-end network trained for 15 epochs
- Best model chosen from validation accuracy
- Same and Cross dataset evaluation
- Adam Optimizer learning rate 0.0002
- GeForce RTX 2060 laptop with GPU and 6GB shared memory+ 16GB total memory

Feature Visualization of MobileNet V2

Output of 32 Filters at Layer 2



Output of 32 Filters of 1280 Filters at Layer 153

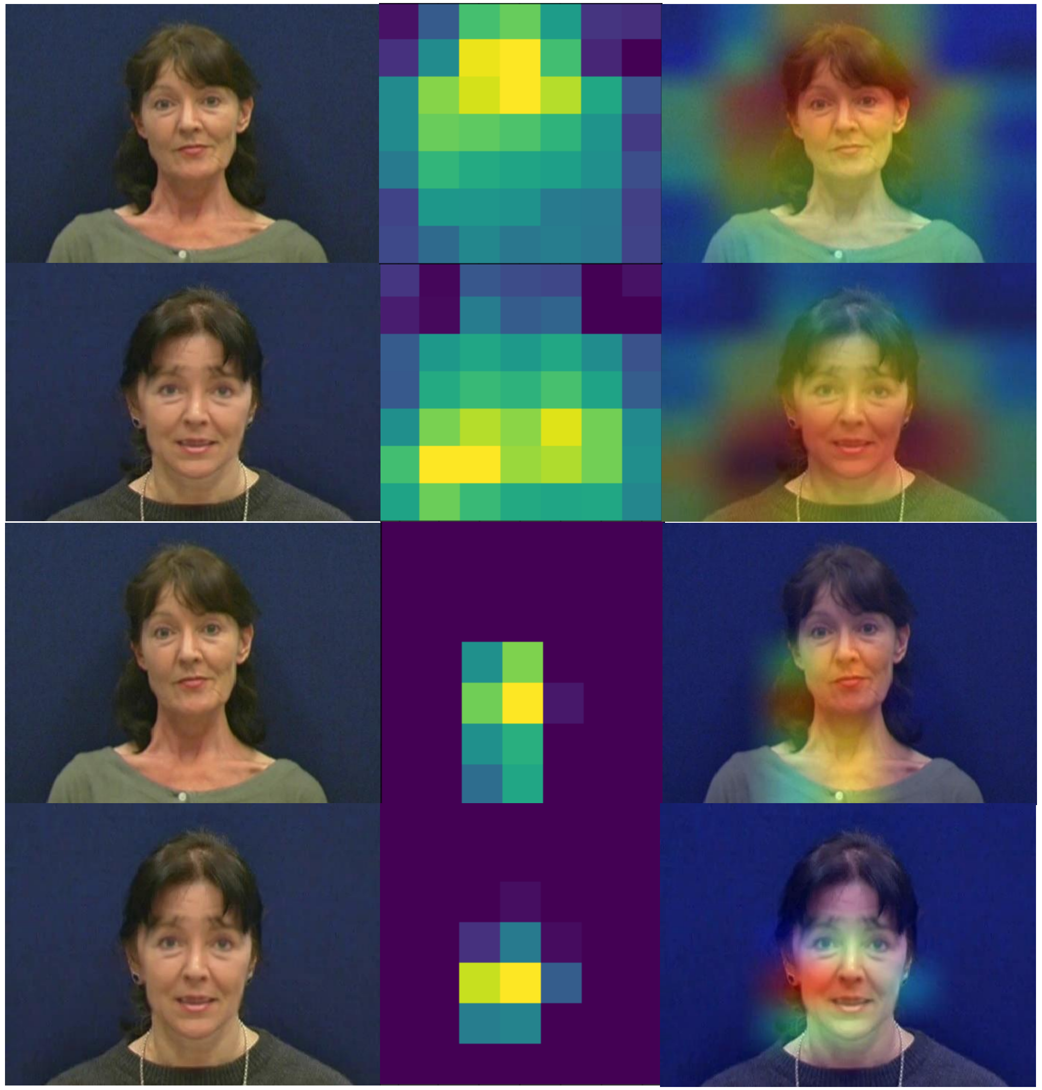


Class Activation Map Visualization (GRAD-CAM)

Predicted **Wrong**

Real

Fake



Pretrained on ImageNet

Predicted **Correct**

Real

Fake

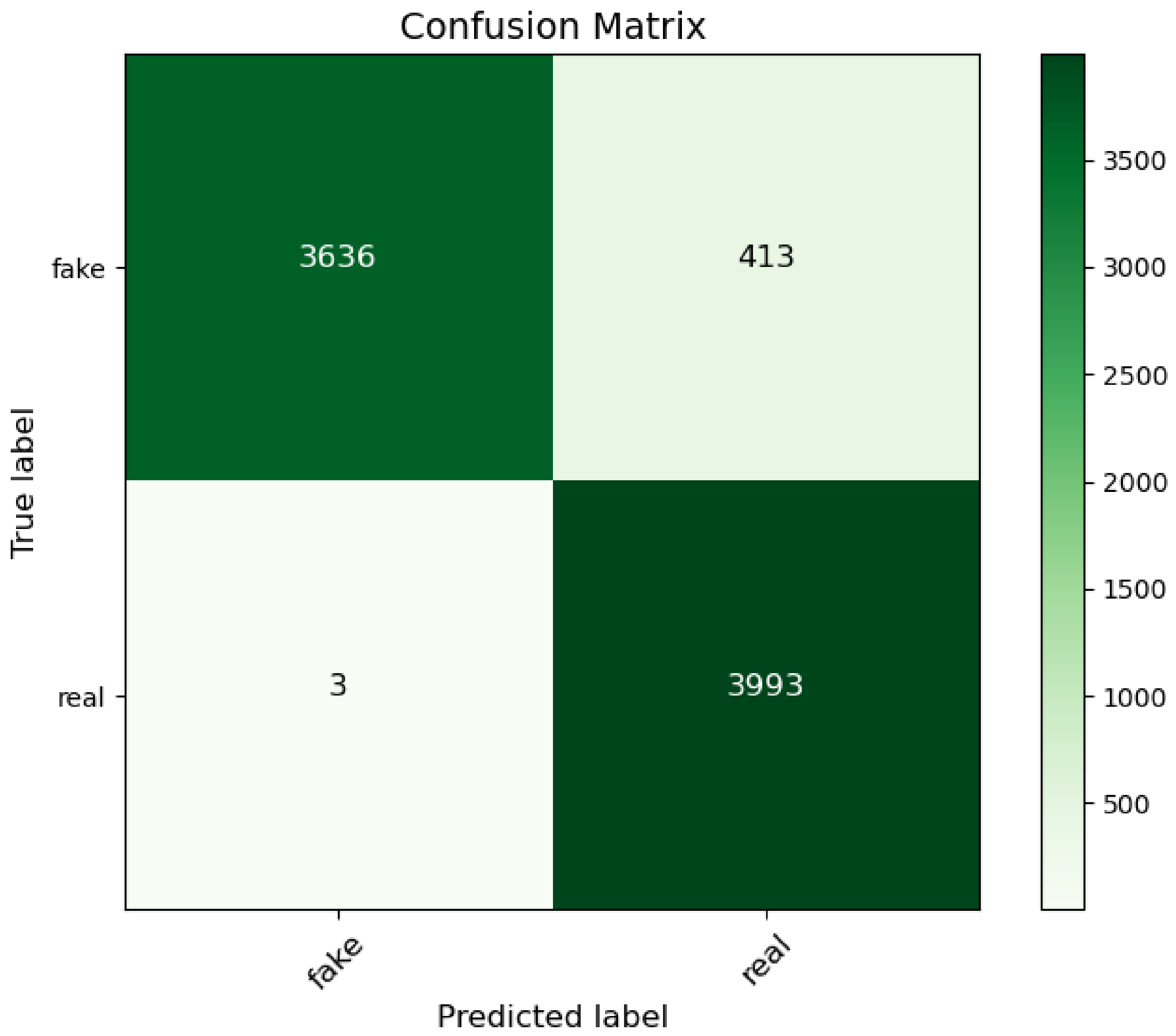
Trained on DF-TIMIT HQ

Accuracy & Inference Time

Training Dataset	Testing Dataset	Accuracy (%)	Inference Time (ms)
DeepfakeTIMIT (HQ)	DeepfakeTIMIT (HQ)	94.83	3.67
DeepfakeTIMIT(LQ)	DeepfakeTIMIT(LQ)	100.00	3.76
DeepfakeTIMIT (HQ)	DeepfakeTIMIT (LQ)	96.91	3.81
DeepfakeTIMIT(LQ)	DeepfakeTIMIT(HQ)	57.38	4.45

For Real images → VidTIMIT dataset

Confusion Matrix



	Predicated Label	
True Label	True Positive (**TP**): Reality : **Fake** Model predicted : **Fake**	False Negative (**FN**): Reality : **Fake** Model predicted : **Real**
	False Positive (**FP**): Reality : **Real** Model predicted : **Fake**	True Negative (**TN**): Reality : **Real** Model predicted : **Real**

Detection Metrices

Test Images	Precision (%)	Recall (%)	F1-score (%)
3,996 Real	100.0	90.0	95.0
4,048 Fake	91.0	100.0	95.0
Macro Average	95.0	95.0	95.0
Weighted Average	95.0	95.0	95.0
Total 8,044	Accuracy (%)	95.0	

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \times 100\%$$

$$Precision = \left(\frac{TP}{TP+FP}\right) \times 100\%$$

$$Recall = \left(\frac{TP}{TP+FN}\right) \times 100\%$$

$$F1-score = \left(\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}\right) \times 100\%$$

Performance Comparison

Papers	PERFORMANCE (%) For DF-TIMIT LQ	PERFORMANCE (%) For DF-TIMIT HQ
Matern et al. [2019]	AUC = 77.00	AUC = 77.30
Yang et al. [2019]	AUC = 55.10	AUC = 53.20
Afchar et al. [2018]	AUC = 87.80	AUC = 68.40
Zhou et al. [2018]	AUC = 83.50	AUC = 73.50
Nguyen et al. [2019]	AUC = 78.40	AUC = 74.40
Proposed Method [2021]	**ACC = 100.00**	**ACC = 94.83**

Conclusions & Future work

- Proposed a CNN based model for Detection of Deep-Morphed Deepfake images in context of Smart City facilities.
- Detected FSGAN generated images
- Light Weight model - makes the Registration Process easy and not localized to Council Office
- High Accuracy
- As future work, generalizability of the model can be obtained

Thank You!!