# IRREGULAR IDENTITY IN DEEPFAKE VIDEOS

# INTRODUCTION

- **Deepfake rise:** Media generation technology enables easy creation of deepfake videos.
- **Key issues:** Defamation, fake news, political manipulation, etc.
- **Current challenges**:
  a. Lack of interpretability in deepfake detection
  b. Poor performance on low-quality videos.

## Core Issue

- Frame-by-frame deepfake generation disrupts facial identity consistency.
- Current Deepfake detectors are limited in explainability and struggle with low-quality Deepfakes

## Objective

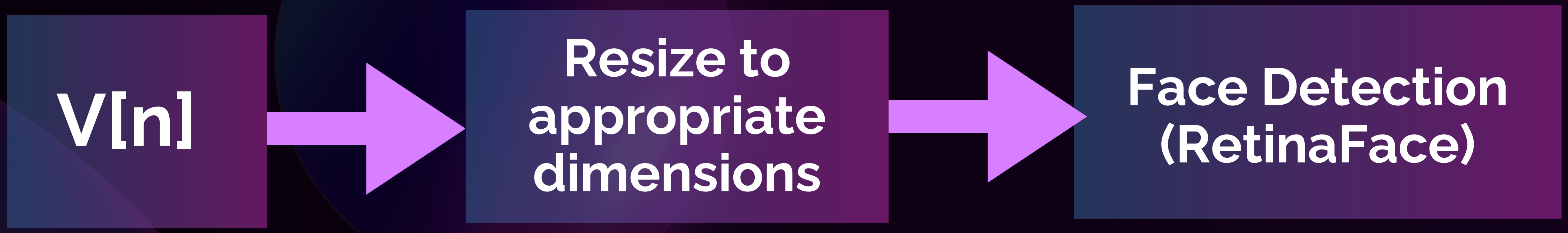
- Develop an interpretable and robust deepfake detection method.
- Focus on identifying irregularities in facial identity features over time.
- Improve detection performance on low-quality videos and ensure generalizability to unseen datasets.

# DATA PREPROCESSING & TRAINING

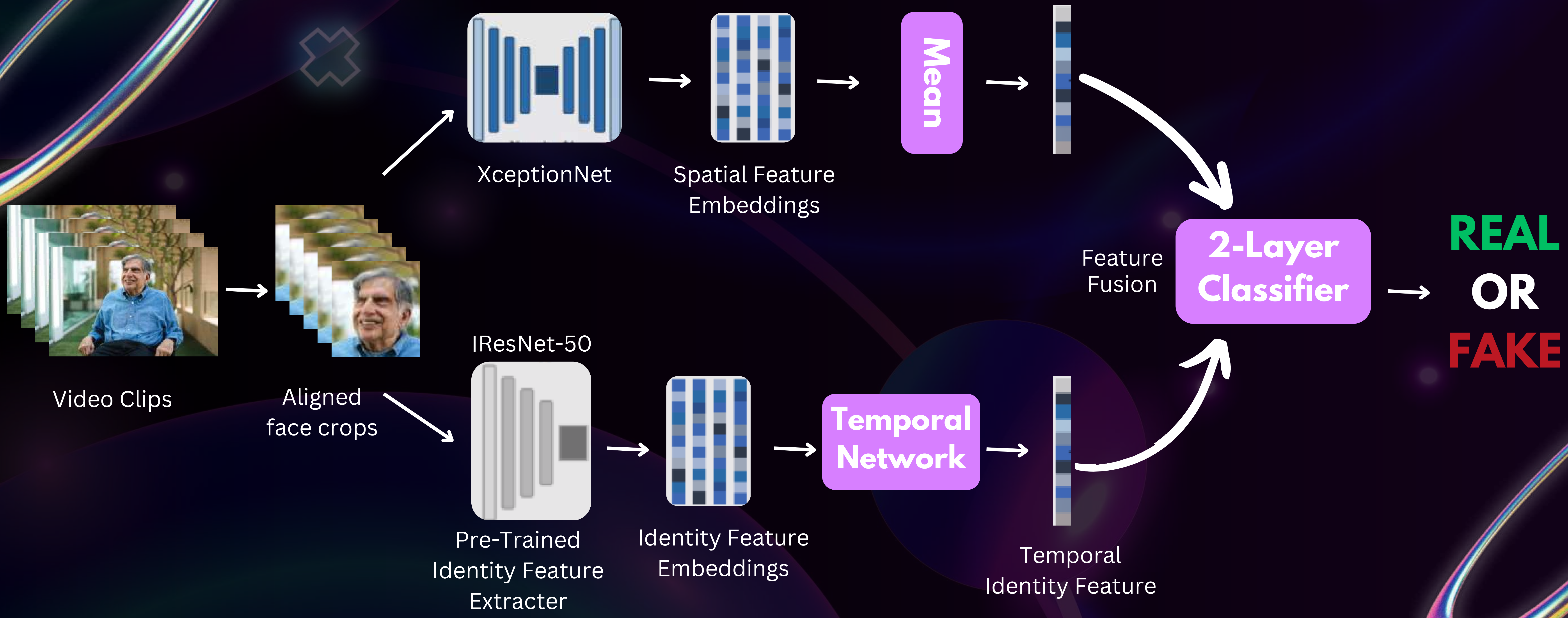
Dataset: FaceForensics++ (FF++) for training, Celeb-DF v2 for cross-dataset evaluation.

Preprocessing: Detect and align faces using RetinaFace.

Faces are cropped and resized to 112x112 for identity feature extraction and 299x299 for spatial feature extraction.


V[n] → Resize to appropriate dimensions → Face Detection (RetinaFace)

# TWO BRANCH ARCHITECTURE


XceptionNet
Spatial Feature Embeddings
Mean
Video Clips
Aligned face crops
IResNet-50
Pre-Trained Identity Feature Extracter
Identity Feature Embeddings
Temporal Network
Temporal Identity Feature
Feature Fusion
2-Layer Classifier
REAL
OR
FAKE

# Spatial Feature Analysis

- Spatial artifacts: Manipulations in deepfakes leave small spatial artifacts that are detectable.
- Feature extraction: XceptionNet is used to extract spatial embeddings f_space[n] for each frame
- Feature aggregation: The spatial embeddings across frames are averaged to generate a summary feature for classification:

$$f_{space}[n]=D_{space}(V[n])$$

$$f_{mean}= 1/N(\sum_{n=0}^{n-1} f_{space}[n])$$

# Identity Feature Analysis

- Identity inconsistency: Deepfakes generate identity features that are less stable across frames.
- Feature extraction: Identity features fid are extracted from each frame using IResNet-50 (pre-trained on face recognition).

$$f_{id}[n]=D_{id}(V[n])$$

where V[n] is the n-th frame and D_id is the IResNet operator.

# Identity Feature Analysis

- Temporal analysis: A temporal model analyzes the series of identity embeddings over time.

$$f_{time} = D_{time}(f_{id}[0], f_{id}[1], \ldots, f_{id}[N-1])$$

where f_time is the temporal embedding representing identity consistency.

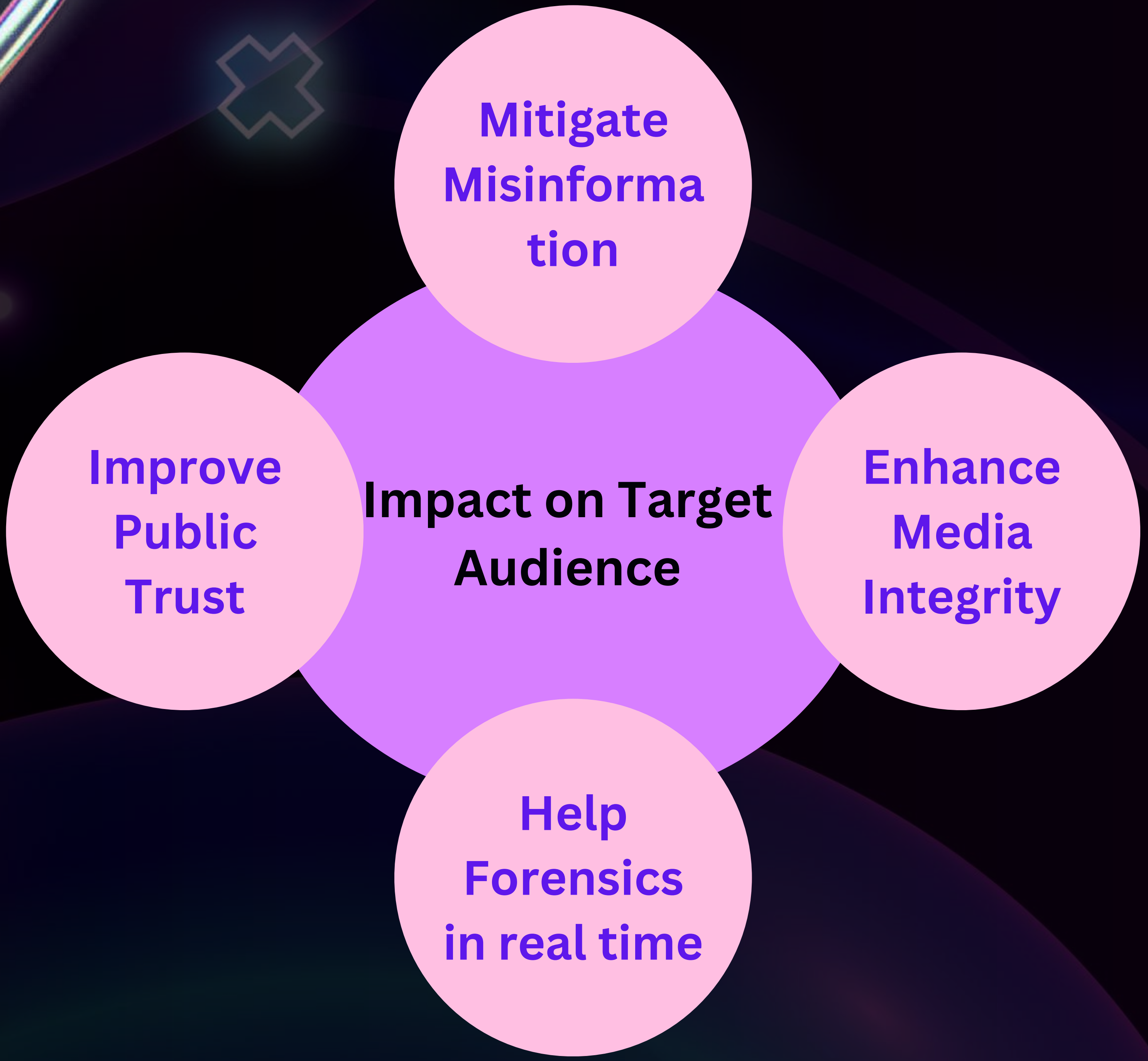
# MODEL ARCHITECTURE

Dual-branch architecture for processing facial identity features and spatial inconsistencies simultaneously.

- **Input:** Cropped, aligned face images.
- **Output:** Real or Fake classification.
- **Components:**
  - Pre-trained identity feature extractor.
  - Temporal network for identity inconsistencies.
  - Spatial feature extractor and final classification layer.

$$\hat{Y} = D_{MLP}(f_{time}, f_{mean})$$

# IMPACT AND BENEFITS



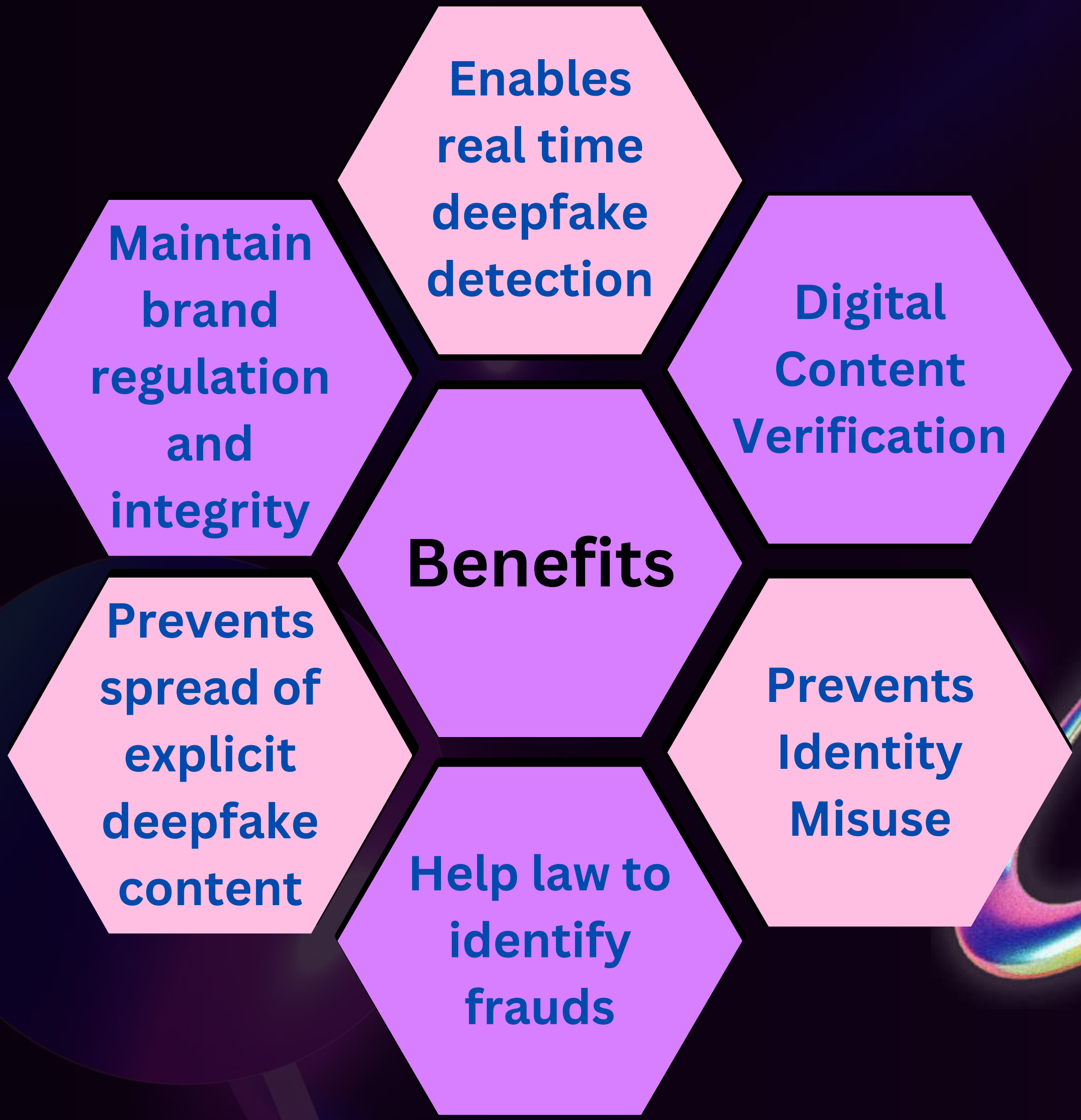
Impact on Target Audience

Mitigate Misinformation

Enhance Media Integrity

Help Forensics in real time

Improve Public Trust



Benefits

Enables real time deepfake detection

Digital Content Verification

Prevents Identity Misuse

Help law to identify frauds

Prevents spread of explicit deepfake content

Maintain brand regulation and integrity

# REFERENCES

- Liu, H., Bestagini, P., et al., "It Wasn’t Me: Irregular Identity in Deepfake Videos," IEEE ICIP 2023.
- R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking DeepFakes with simple Features," *arXiv preprint arXiv:1911.00686v3*, 2020.
- Dataset : FaceForensics++ and Celeb-DF v2.

# THANK YOU!