Deepfakes Detection (Videos)

- A battle for the truth.

❑ **Group project presentation by:**

- **Vivek Kumar Dutta [35000122030]**
- **Sanchita Tewary [35000122020]**
- **Souvik Sen [35000122023]**

❑ **Guided by Mentor :**

- **Prof. Pabitra Roy**

Introduction – What's a deepfake ?

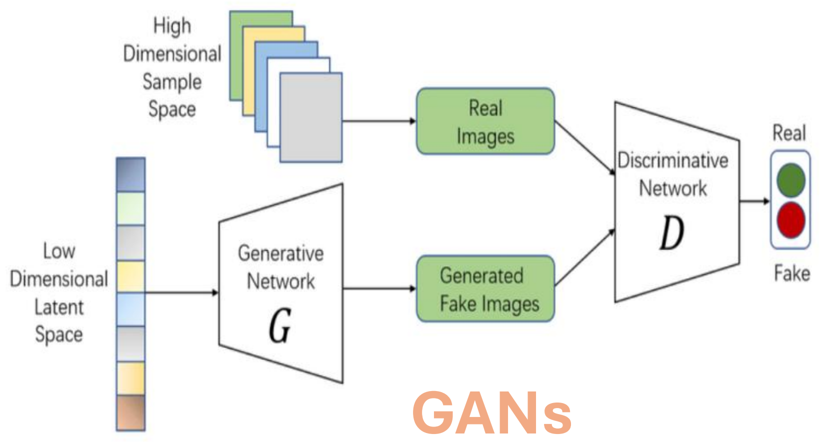
A piece of **synthetic media** (like a video, image, or audio recording) that has been convincingly **manipulated** or entirely **fabricated** using **artificial intelligence (AI) and deep learning techniques**.

Original Video

Fake Video

A Deepfake Video at its simplest is **Swapping the Face** of a real person in a video with the face of a different person , either real or AI Generated.

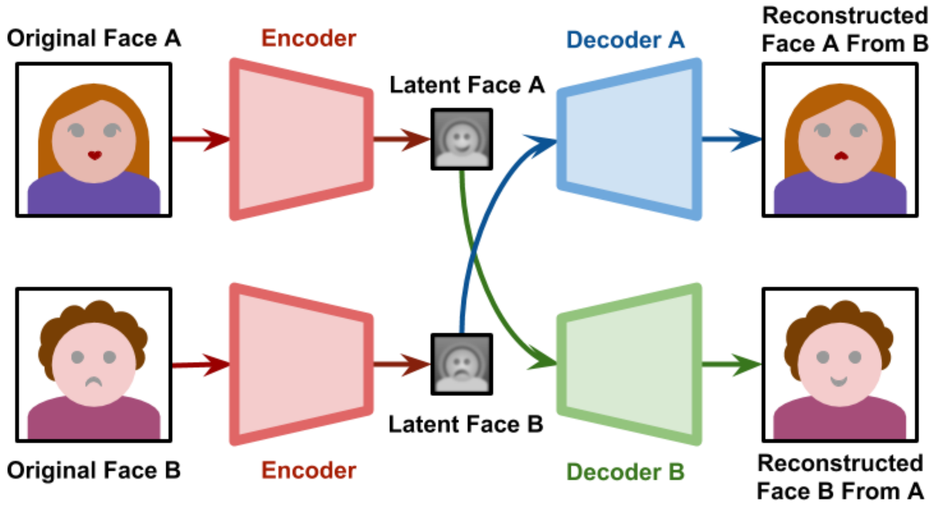
And **How** is that done ?



GANs

powerful deep learning models using two competing neural networks—a **Generator** and a **Discriminator**—to create new, realistic data that mimics a training dataset, with the Generator trying to fool the Discriminator, and the Discriminator trying to spot fakes, leading to continuous improvement in data generation quality

Unsupervised neural networks that learn efficient data representations by **compressing input** into a lower-dimensional "latent space" (encoding) and then **reconstructing** the original data from that compressed form (decoding)

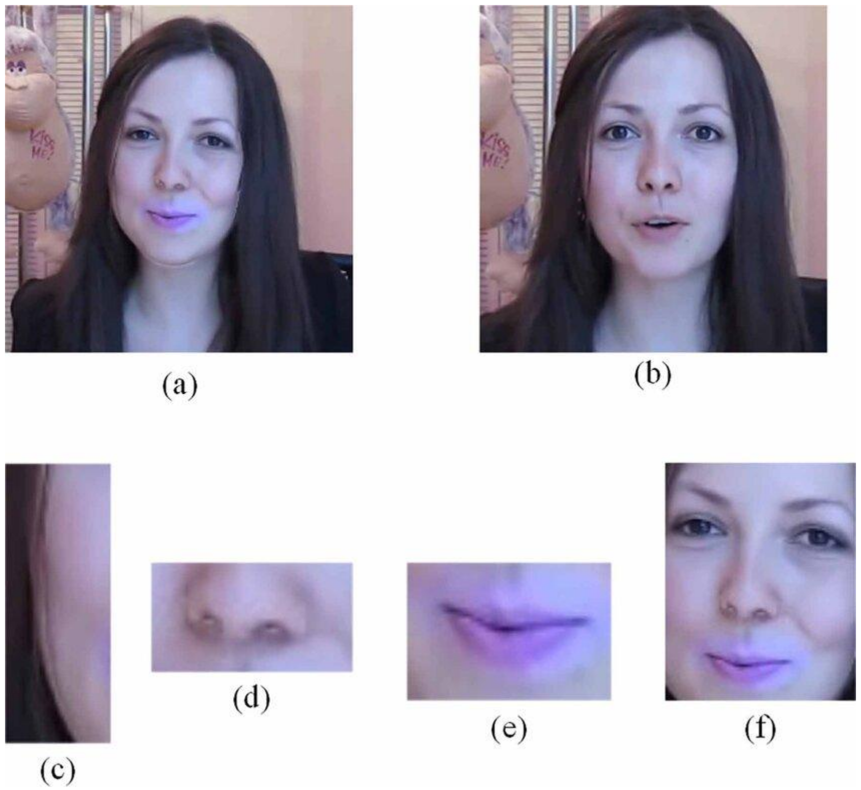


Autoencoders

Alright, so that's how they are created.

Now How can we teach the computer to differentiate between a Real Video and a Fake Video ?

Spatial (Frame-level) Artifact Detection



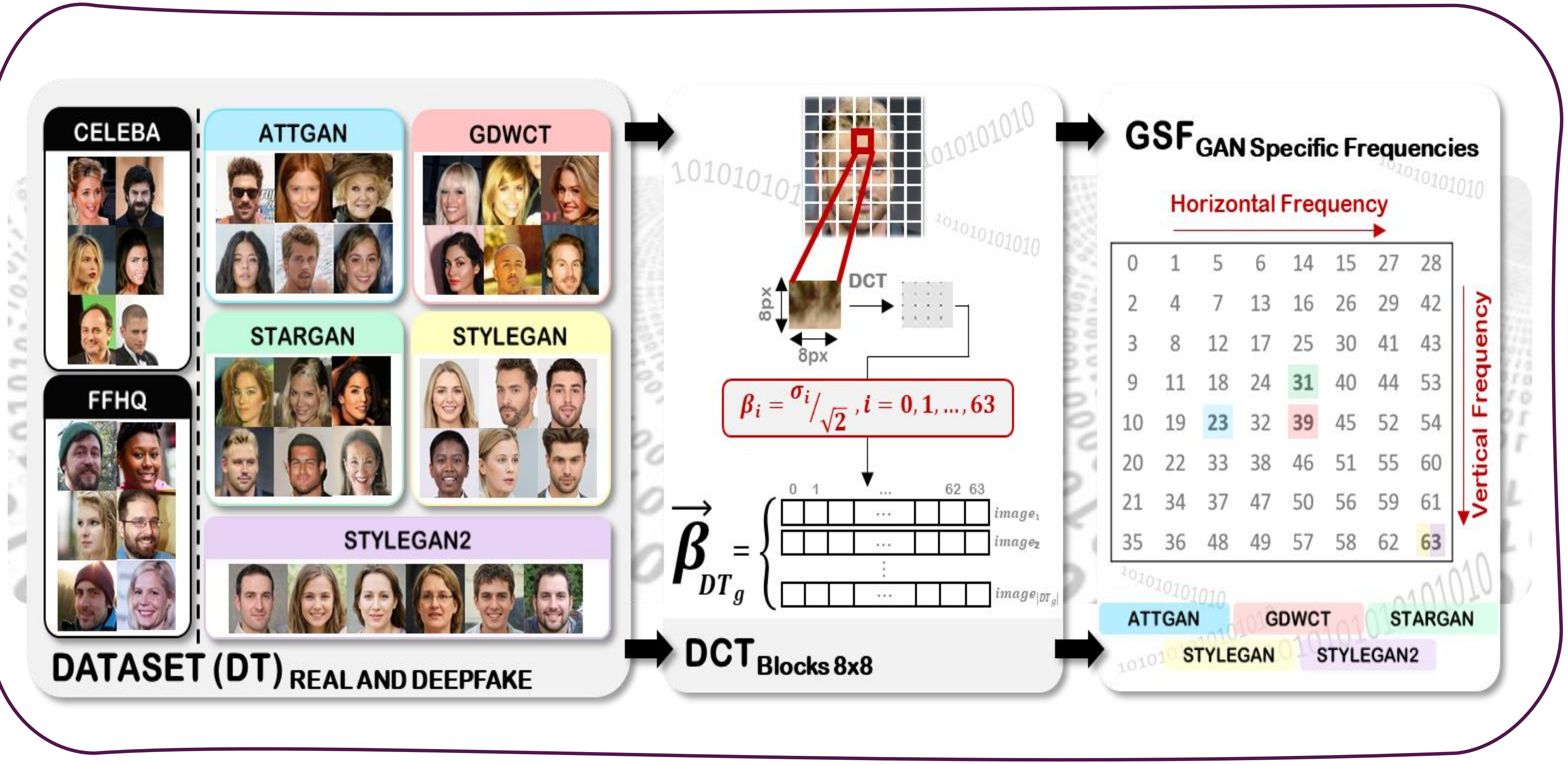
Deepfake generation models often struggle with **fine-grained facial** details such as **skin texture**, **eye boundaries**, **teeth**, **hairlines**, and **face–background blending.**

CNN based Neural Networks excel in finding these **hidden spatial artifacts** very efficiently.

However there are limitations – This method ignores motion, i.e. the **Time Dimension**. making them vulnerable to high-quality fakes that are visually clean but temporally inconsistent

Frequency-domain and Compression Artifact Analysis

Deepfake generators often leave **telltale traces** in high-frequency components due to **up-sampling, convolutional kernels**, and **GAN training dynamics.** These artifacts may be invisible spatially but become evident after **transforms** like DCT, FFT, or wavelet decomposition.



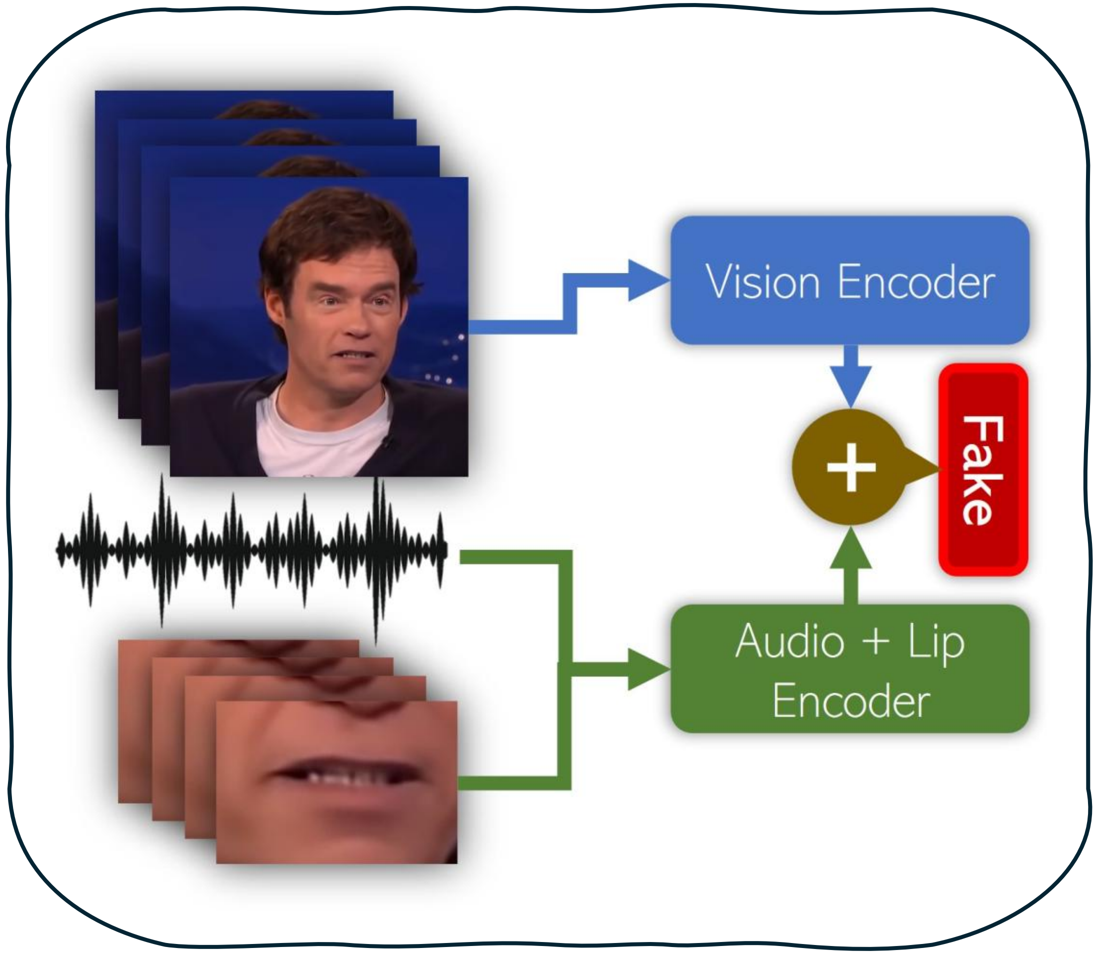
Temporal and Motion based Detection

Allows the model to find **Sequential Evidences** while analyzing the **Time dimension** !

(a) Local-consecutive short-term inconsistency (intra-group)

(b) Long-term inconsistency (inter-group)

Audio-Visual Consistency Checking

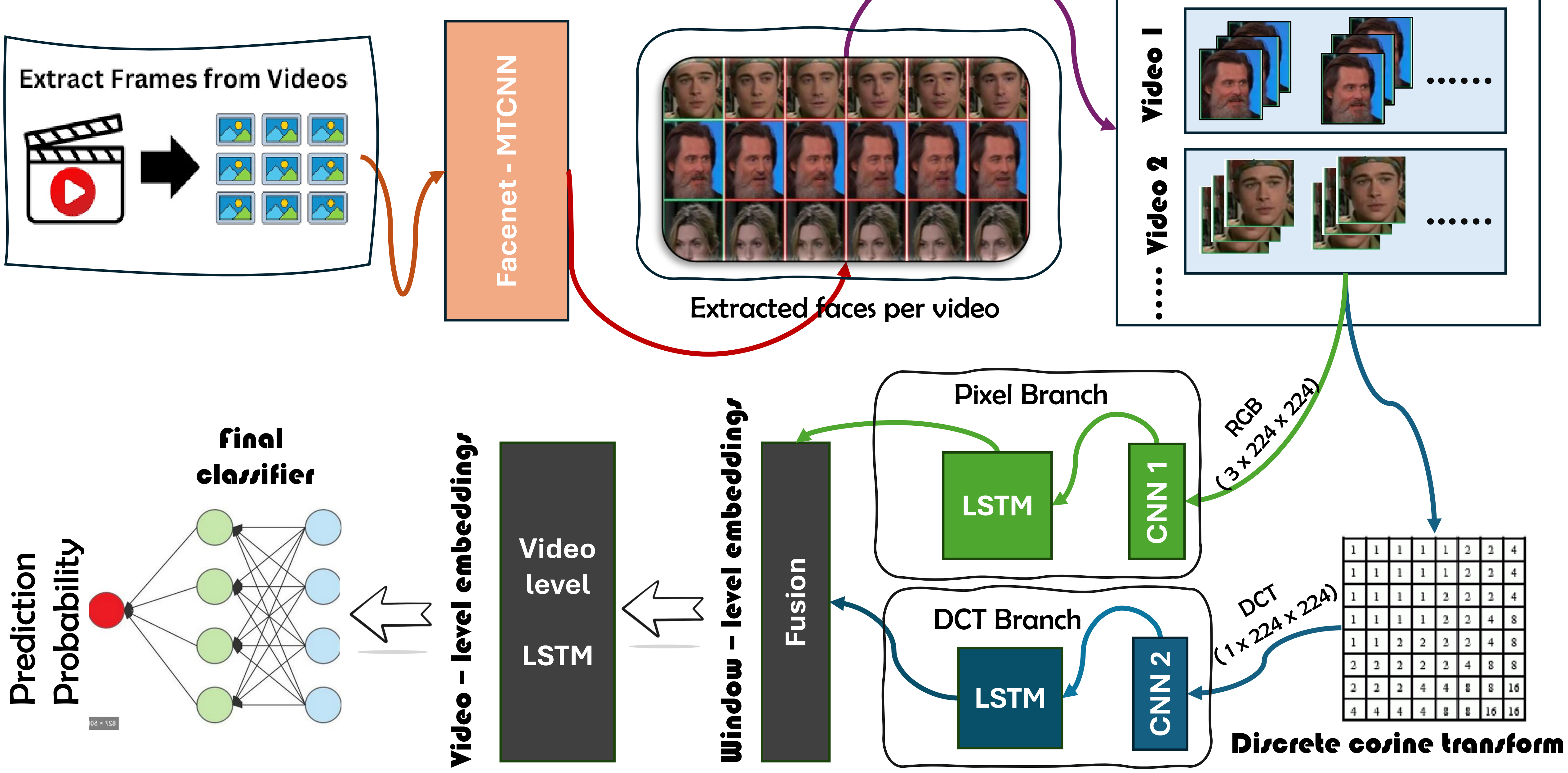


In many deepfake videos, the face and voice are generated or modified separately. This opens the door to **cross-modal inconsistencies**.

Audio–visual methods check whether lip movements match the spoken phonemes, whether facial expressions align with vocal emotion, or whether **head motion correlates with speech dynamics.**

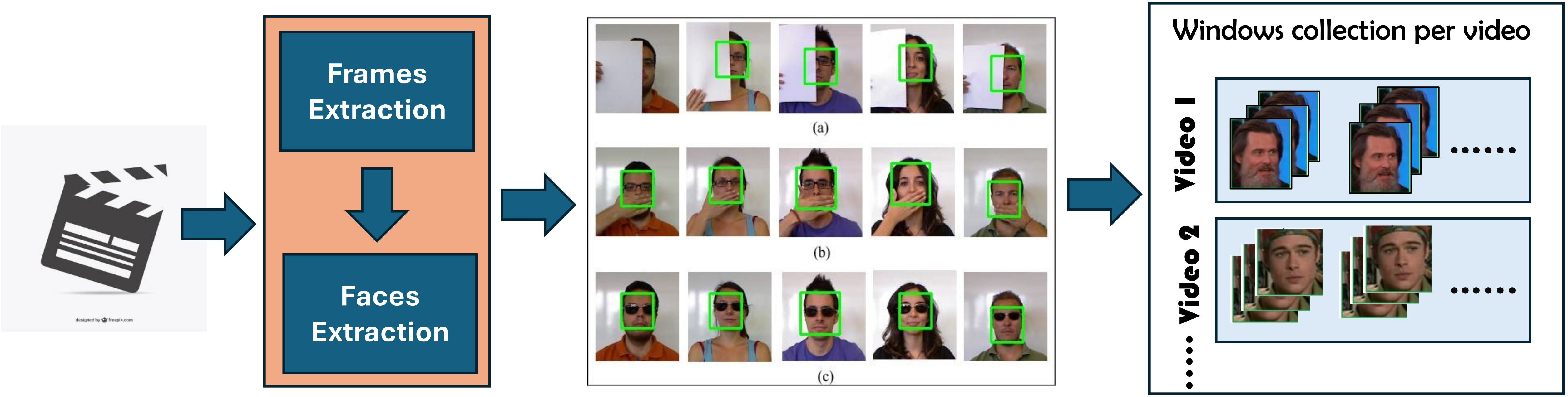
These systems jointly **embed** audio and video streams and look for mismatches in their temporal alignment. Even when both modalities are realistic independently, their relationship can expose manipulation.

Our Model Structure



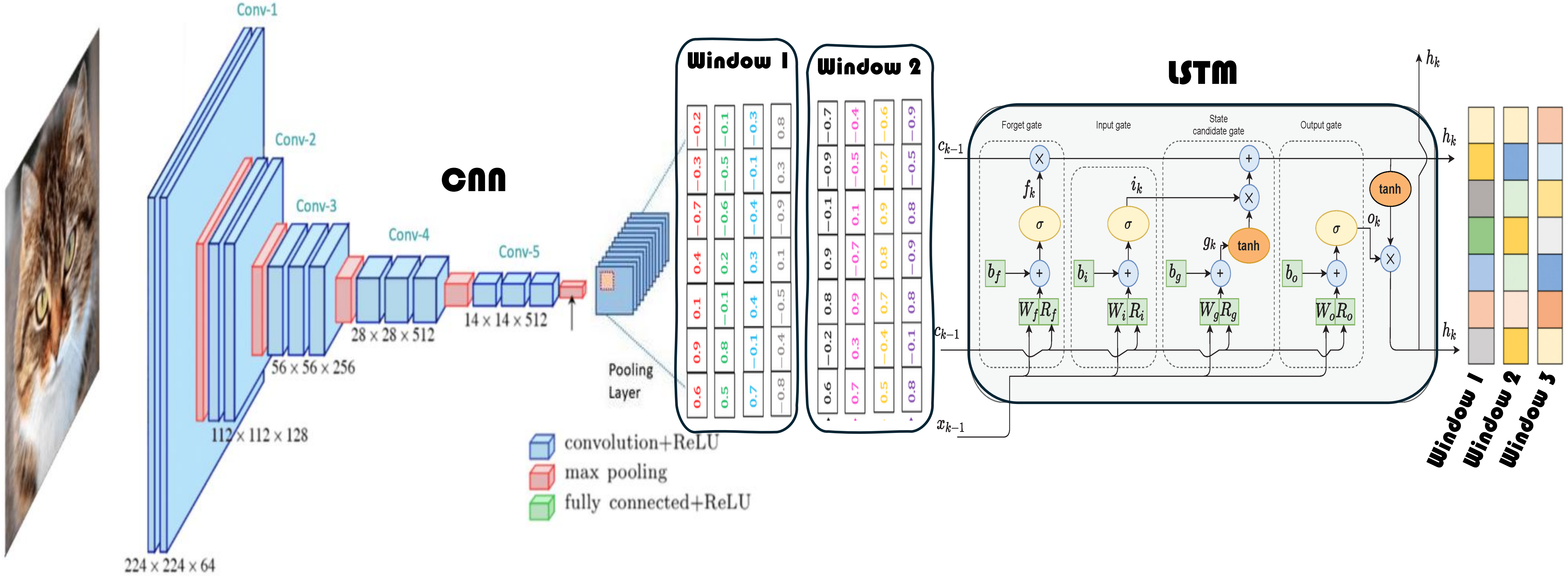
Level 0: Faces **Extraction** and **Data** Preparations

This phase Deals with **Preparing the Data** to feed to the Model. It involves **Extraction** of individual **faces** from individual **frames** across the video. Then, Consecutive sampled frames are grouped into '**Windows**'. Each video is represented as a **sequence** of **Multiple** Windows.



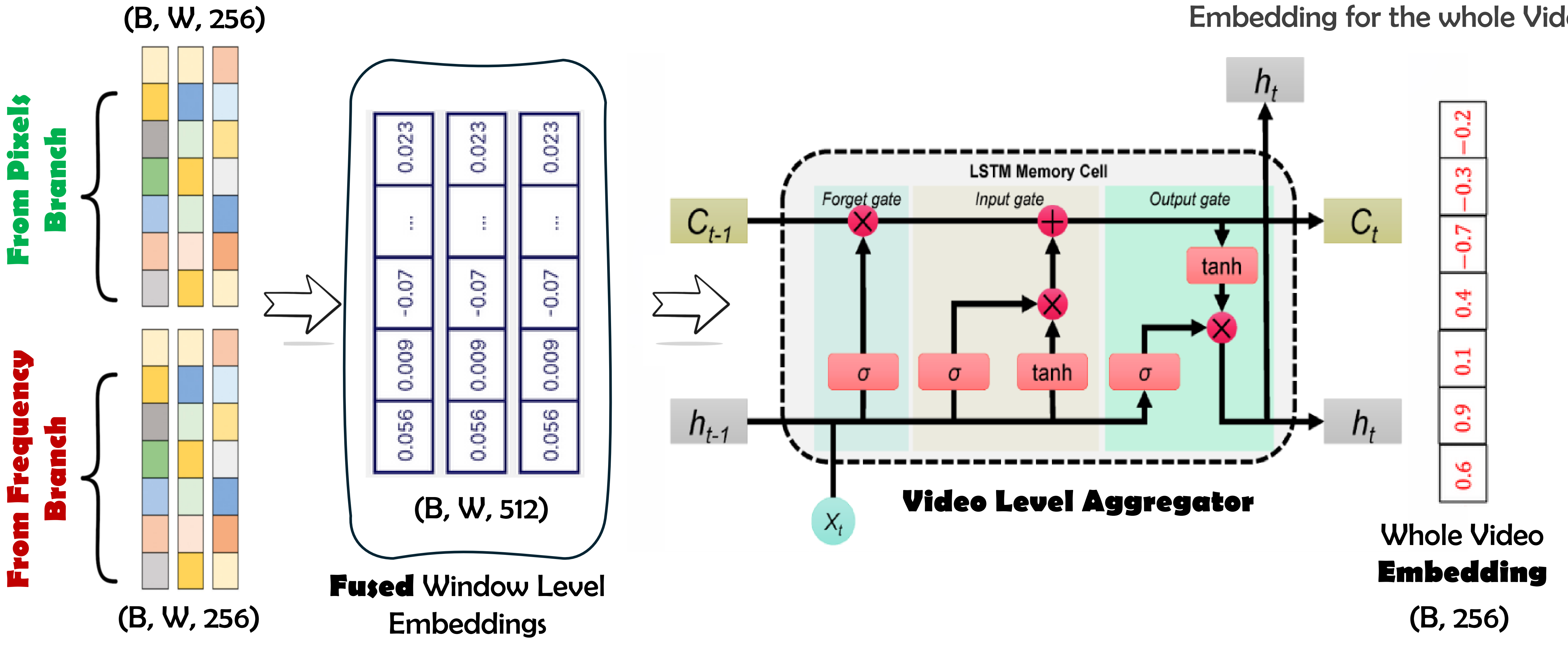
Level 1 : Window Level Aggregations

A window is a collection of consecutive frames in a video. Each frame is passed onto the **CNNs** to produce the **frame- level embeddings**. Later they are sequentially passed onto **LSTM** to produce **Window-level embeddings**.



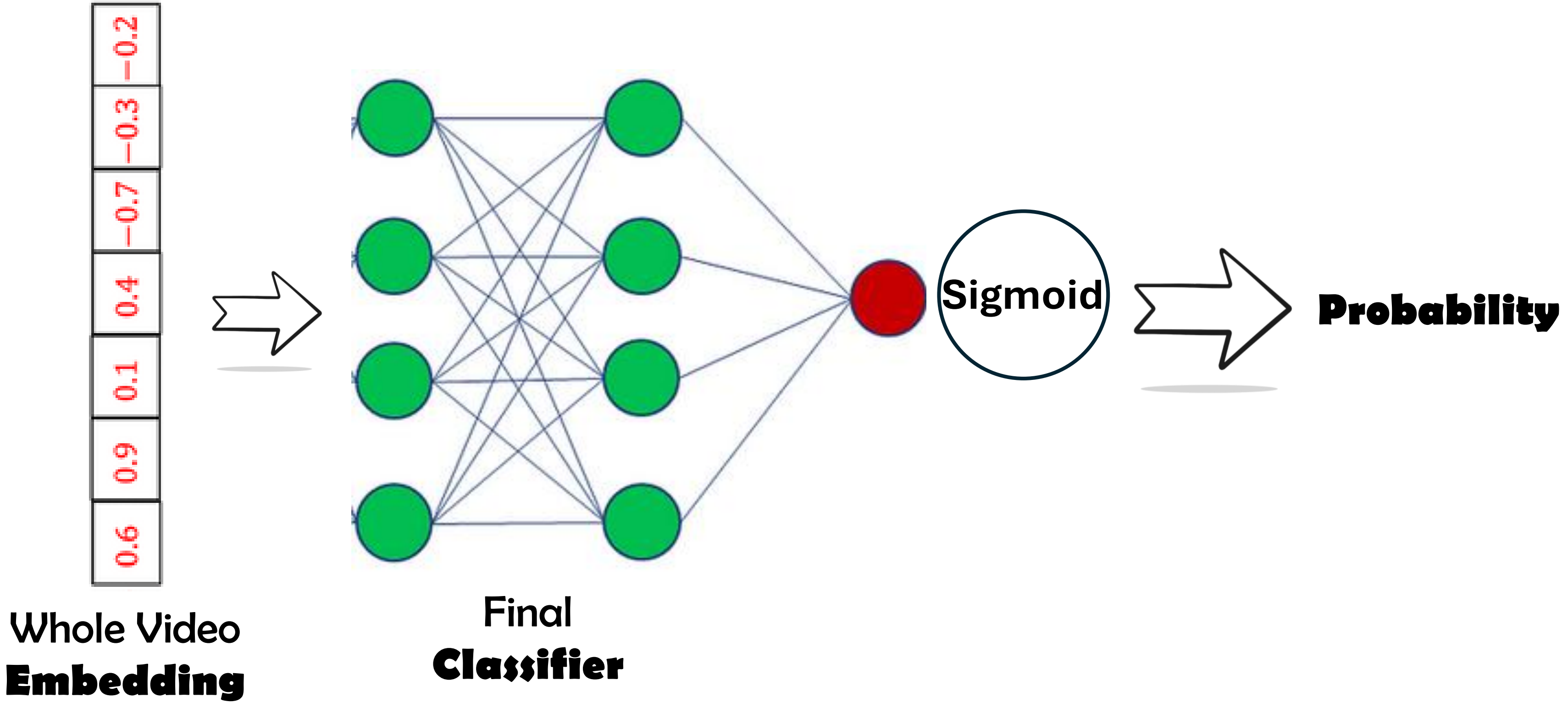
Level 2: Video Level Aggregation

The Outputs from both the **'Pixel-branch'** as well as the **'Frequency branch'** (aka the DCT branch) are fused together to produce a joint Embedding vector representing the entire Window. These are then fed to the LSTM to produce A Single Embedding for the whole Video



Level 3: Classification

The Final Embedding vector is fed to a Classifier and the Probability is estimated.

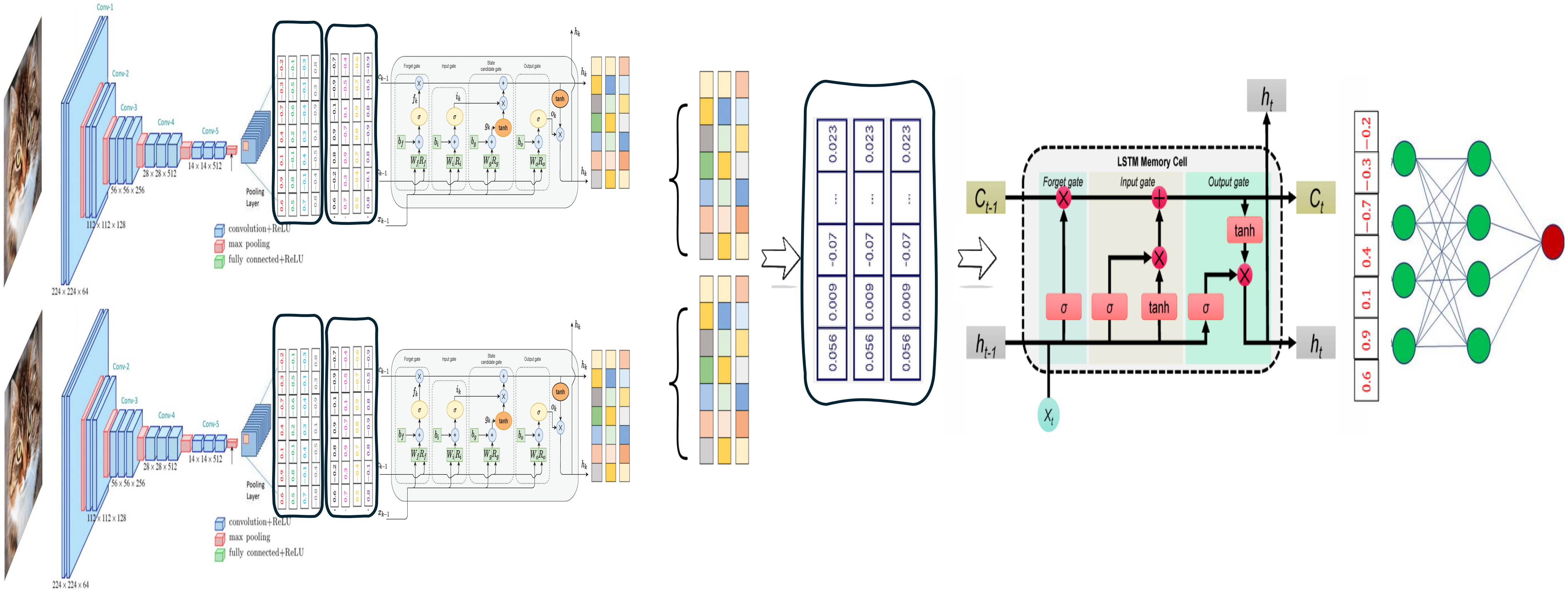


Training the Model : Hierarchical Learning

Training all the Models at once causes **Significant Overfitting** of the Model.

Phase 1 training

Phase 2 training



Current Limitations of the Model

1. Single-Dataset Training Constraint

The proposed deepfake detection model is trained and evaluated exclusively on the **Celeb-DF v2** dataset. While this dataset is widely used and contains high-quality deepfake videos, training on a single dataset limits the model's ability to generalize across different real-world conditions. Deepfake generation techniques, video resolutions, compression levels, and post-processing pipelines vary significantly across datasets and platforms. As a result, the learned spatial, frequency, and temporal patterns may be biased toward Celeb-DF-specific artifacts. This can lead to a noticeable performance drop when the model is tested on unseen datasets such as FaceForensics++, DFDC, or real social-media videos, highlighting a current limitation in cross-dataset robustness.

Computational Cost and Modal Limitations

The hierarchical nature of the proposed model—combining face extraction, dual CNN branches, window-level LSTMs, and a video-level LSTM—introduces significant computational overhead. This makes real-time deployment challenging, especially on resource-constrained devices. Additionally, the model relies solely on visual information and does not incorporate **audio-visual consistency checks**, which can be highly informative in many deepfake scenarios where audio and video are generated separately. Furthermore, the model assumes reliable face detection in every frame; failures in face detection due to occlusion, extreme poses, or low video quality may negatively impact performance. Addressing these limitations is an important direction for future improvement.

Thank You !