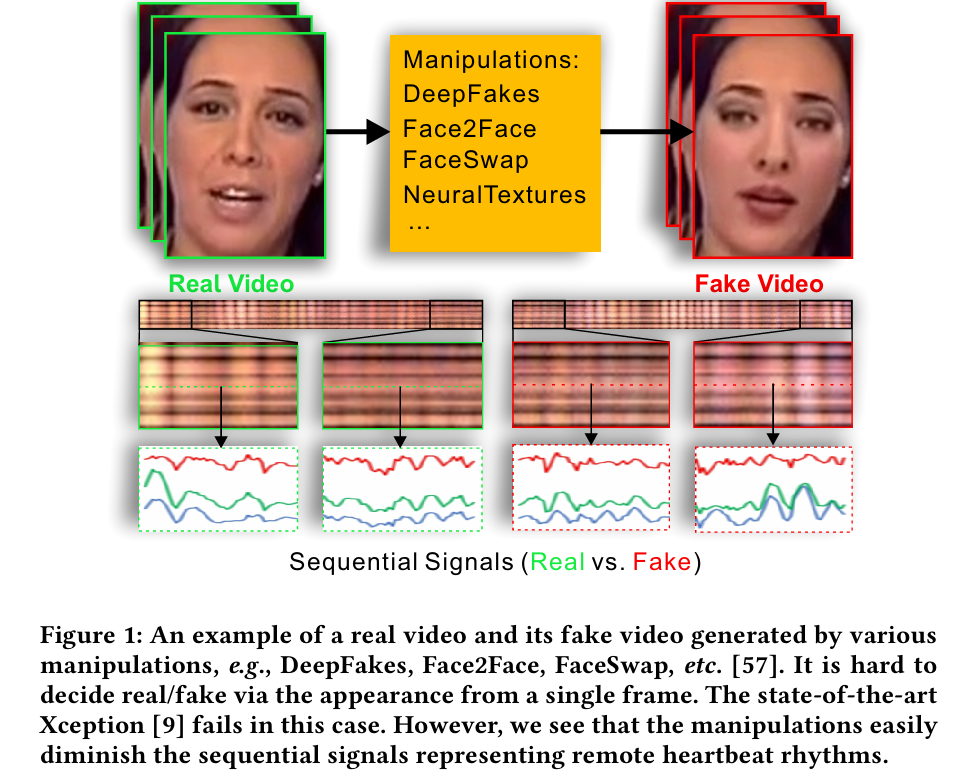
Our method is designed to detect DeepFake videos by analyzing heartbeat rhythms from facial regions using **remote photoplethysmography (rPPG)**. The methodology follows a systematic flow:



**1. Dataset Preparation**

**Step 5.1: Datasets Available**

* **FaceForensics++ (FF++)**: Contains DeepFake, Face2Face, FaceSwap, NeuralTextures and FaceShifter videos.
* **Out of them, we used Real and Deepfake videos for training our model, and evaluated it on Real and** Face2Face videos

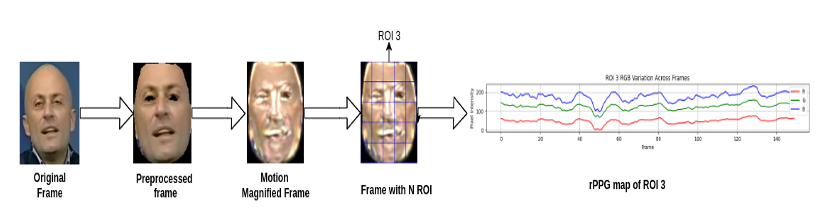
**2. Data Preprocessing and Face Segmentation**

**Step 1.1: Face Detection and Landmark Extraction**

* **Face Detection:** Detect the face in each frame using **MTCNN** (Multi-task Cascaded Convolutional Networks).
* **Facial Landmarks:** Identify **81 key landmarks** on the face using **Dlib**.
* **Region of Interest (ROI) Selection:**
  + Remove **eyes and background** as they introduce noise.
  + Focus on **forehead, cheeks, and under-eye areas**, where heartbeat signals are strongest.

**Step 1.2: Face Tracking and Stabilization**

* If multiple faces are detected, retain the one that is closest to the previously detected face.
* Frames without detected faces are discarded (if more than 50 frames are lost, the video is skipped).



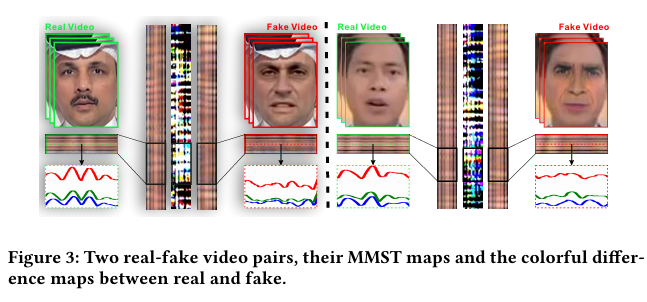
**3. Motion-Magnified Spatial-Temporal Representation (MMSTR)**

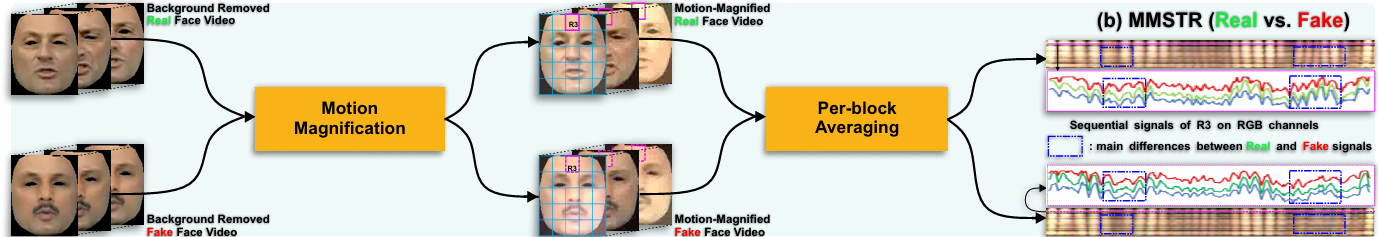
**Step 2.1: Motion Magnification for Heartbeat Enhancement**

* **Why?** The heartbeat signal in facial regions is subtle and may not be directly visible.
* **How?** Apply the **Eulerian Video Magnification (EVM)** technique to amplify small color changes caused by blood flow.
* **Output:** A **motion-magnified face video** where color changes due to the heartbeat are enhanced.

**Step 2.2: Generating the MMST Map**

* Divide the **face into N non-overlapping blocks (ROIs)** (e.g., 5x5 grid = 25 blocks).
* Extract the **average RGB color intensity per block** over time.
* Construct an **MMST (Motion-Magnified Spatial-Temporal) Map**, where:
  + **Rows** = different facial regions (N blocks)
  + **Columns** = time (frames)
  + **Values** = RGB intensity variations (representing heartbeat patterns)





**4. Dual-Spatial-Temporal Attentional Network (Dual-ST AttenNet)**

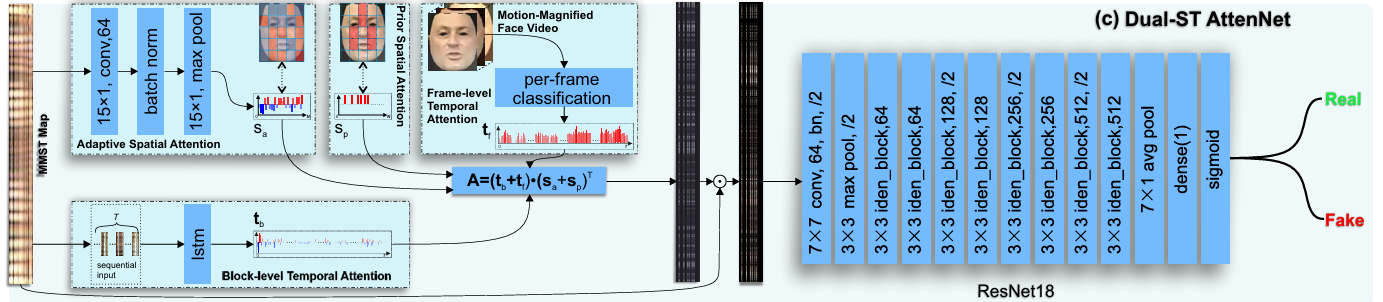
DeepRhythm utilizes a **dual-attention mechanism** to focus on meaningful areas while ignoring noise.

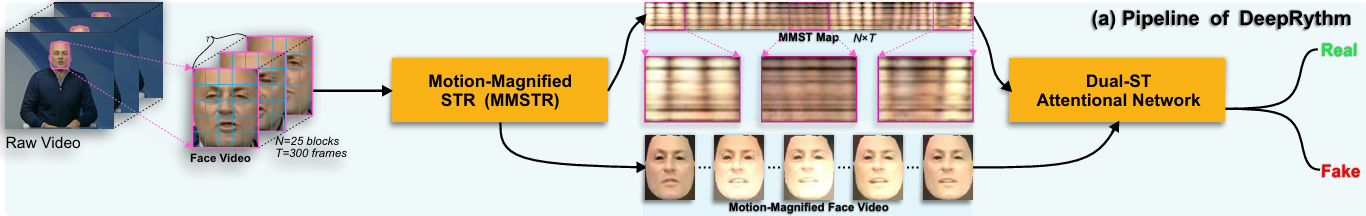
**Step 3.1: Spatial Attention (Where to Focus?)**

* Some facial regions provide **stronger heartbeat signals** than others.
* The model applies a **Dual-Spatial Attention Mechanism**:
  1. **Prior Spatial Attention (Fixed Weights):** Focuses on pre-defined robust regions (e.g., under the eyes).
  2. **Adaptive Spatial Attention (Learned Weights):** Adjusts dynamically based on video conditions (e.g., lighting changes).

**Step 3.2: Temporal Attention (When to Focus?)**

* Some frames contain more **distinctive DeepFake artifacts** than others.
* **Two types of temporal attention are applied:**
  1. **Block-Level Temporal Attention:** Uses **LSTM** to analyze variations in facial regions over time.
  2. **Frame-Level Temporal Attention:** Uses **MesoNet** (a CNN-based model) to assign importance scores to frames.
* The final weight matrix **A = (t \* s⊤) ⊙ X** ensures that the model gives higher importance to frames and regions with strong heartbeat signals.





**5. Deep Neural Network for DeepFake Classification**

* The **weighted MMST map** (with spatial and temporal attention applied) is passed to a **deep neural network** for classification.
* **ResNet18** is used as the final classifier.
* **Adam Optimizer is used**
* The network is trained to **output 1 for fake videos and 0 for real videos**.

