1 Proposed Methodology

We propose a unified cross-modal frequency-aware architecture for robust activity recognition by jointly leveraging RGB, Depth, and IMU modalities. The pipeline involves input preprocessing, hierarchical multi-scale feature extraction, wavelet-based decomposition, attention-based modulation, and classification.

Given a batch size B and temporal window size T, the input comprises RGB video frames $\mathbf{X}_{\text{rgb}} \in \mathbb{R}^{B\times T\times H\times W\times 3}$, Depth video frames $\mathbf{X}_{\text{depth}} \in \mathbb{R}^{B\times T\times H\times W\times 1}$, and IMU time-series data $\mathbf{X}_{\text{imu}} \in \mathbb{R}^{B\times T\times 6}$. These are first transformed into frequency-based 2D representations using domain-specific techniques such as Continuous Wavelet Transform (CWT), HHA encoding, or spectrograms:

$$\begin{aligned} \mathbf{X}_{\text{rgb}}^{\text{2D}} &\in \mathbb{R}^{B\times C_r\times H\times W}, \\ \mathbf{X}_{\text{depth}}^{\text{2D}} &\in \mathbb{R}^{B\times C_d\times H\times W}, \\ \mathbf{X}_{\text{imu}}^{\text{2D}} &\in \mathbb{R}^{B\times C_i\times H\times W}. \end{aligned} \tag{1}$$

A shared Hierarchical Multiscale Convolutional Network (HMCN) is then used to extract modality-specific features. The HMCN operates by first splitting the input channels into s groups $\phi_l \in \mathbb{R}^{B\times C'\times H\times W}$, where $C' = \lceil C/s \rceil$. Each group is processed by depthwise separable convolutions comprising depthwise and pointwise operations:

$$Y_d(i,j,c) = \sum_{m=0}^{k-1}\sum_{n=0}^{k-1} X(i+m, j+n, c)\cdot W_d(m,n,c), \tag{2}$$

$$Y(i,j,k) = \sum_{c=1}^{C'} Y_d(i,j,c)\cdot W_p(c,k) + b_k. \tag{3}$$

These groups propagate features hierarchically in two directions. The left wing processes as:

$$Y_l^{\text{left}} = \begin{cases} \text{Conv}_{\text{dwp}}(\phi_1), & l = 1 \\ \text{Conv}_{\text{dwp}}(Y_{l-1}^{\text{left}} \oplus \phi_l), & l > 1, \end{cases} \tag{4}$$

and the right wing propagates similarly in reverse:

$$Y_l^{\text{right}} = \begin{cases} \text{Conv}_{\text{dwp}}(\phi_s), & l = s \\ \text{Conv}_{\text{dwp}}(Y_{l+1}^{\text{right}} \oplus \phi_l), & l < s. \end{cases} \tag{5}$$

All outputs are concatenated to form the final feature map:

$$F = \text{Concat}(\{Y_l^{\text{left}}\}, \{Y_l^{\text{right}}\}) \in \mathbb{R}^{B\times C'\times H'\times W'}. \tag{6}$$

This operation is applied independently for each modality, i.e.,

$$F_m = \text{HMCN}(\mathbf{X}_m^{\text{2D}}), \quad m \in \{\text{rgb}, \text{depth}, \text{imu}\}. \tag{7}$$

Each feature map F_m undergoes a 2D Discrete Wavelet Transform (DWT) to produce four subbands:

$$\text{DWT}(F_m) = \{LL_m, LH_m, HL_m, HH_m\}, \tag{8}$$

where each component is of shape $\mathbb{R}^{B\times C'\times \frac{H'}{2}\times \frac{W'}{2}}$.

We then generate a frequency attention gate using RGB subbands:

$$F_{\text{wave}}^{\text{rgb}} = LL_{\text{rgb}} + LH_{\text{rgb}} + HL_{\text{rgb}} + HH_{\text{rgb}}, \tag{9}$$

$$g = \text{GAP}(F_{\text{wave}}^{\text{rgb}}) \in \mathbb{R}^{B\times C'}, \tag{10}$$

$$\alpha = \sigma(W_2(\text{ReLU}(W_1(g)))) \in \mathbb{R}^{B\times C'}. \tag{11}$$

This attention vector α is used to modulate the subbands of Depth and IMU features via channel-wise multiplication:

$$\begin{aligned} LL'_m &= LL_m \odot \alpha, \quad LH'_m = LH_m \odot \alpha, \\ HL'_m &= HL_m \odot \alpha, \quad HH'_m = HH_m \odot \alpha. \end{aligned} \tag{12}$$

The modulated subbands are recombined using inverse DWT to obtain refined feature maps:

$$F_m^{\text{mod}} = \text{IDWT}(LL'_m, LH'_m, HL'_m, HH'_m), \quad m \in \{\text{depth}, \text{imu}\}. \tag{13}$$

Finally, all features are concatenated and passed through global average pooling followed by a fully connected layer and softmax for classification:

$$F_{\text{fused}} = \text{Concat}(F_{\text{rgb}}, F_{\text{depth}}^{\text{mod}}, F_{\text{imu}}^{\text{mod}}), \tag{14}$$

$$\hat{y} = \text{Softmax}(W_{\text{cls}}(\text{GAP}(F_{\text{fused}}))). \tag{15}$$

The output $\hat{y} \in \mathbb{R}^{B \times \text{num\_classes}}$ represents the predicted activity labels for each sample in the batch.