## Pose Estimation CNN Architecture

Figure: This architecture employs an encoder-decoder design, leveraging information from multiple views to enhance performance. A shared encoder processes each input view individually, generating an encoded representation for each view. These individual encodings are concatenated into a unified feature map, which is subsequently fed into a shared decoder along with the individual

encoded view. The decoder produces C heatmap channels, where each channel represents the probability density of feature points at each pixel. The feature point for each heatmap is determined as the pixel with the highest intensity value. The concatenated representation for each view is given by:  $Decoder-Input_i = [E_i, E_1, E_2, E_3, E_4],$ where  $E_i$  is the encoded representation of the specific view, and  $E_1, E_2, E_3, E_4$  are the encodings of all views. Encoder output 1 128 128 128 256 256Encoder output 2 128 128 128 128 128 128 64 64 64256 256Encoder output 3 1 128 128 128 128 128 128 Encoder output 4 128 128 128  $128 \ 128 \ 128$