

ABSTRACT

NEURAL NETWORK SYSTEM FOR 3D POSE ESTIMATION

A system for 3D pose estimation of people is described, comprising an encoder neural network configured to extract multi-view image features from 2D input images obtained from different camera views, and a decoder neural network configured to receive the multi-view image features and predict 3D joint locations of the people in the 2D input images. The decoder neural network comprising a projective-attention mechanism configured to: determine a 2D projection of a predicted 3D joint location for the camera views and assign them as anchor points; apply an adaptive deformable sampling strategy to gather localized context information of the camera views and to learn deformable offsets, and based on the deformable offsets, determine the deformable points for the anchor points; generate attention weights based on the multi-view image features at the anchor points; and apply the attention weights to aggregate the multi-view image features at the deformable points.

[Figure 3]