Denoising Diffusion for 3D Hand Pose Estimation from Images

Maksym Ivashechkin, Oscar Mendez, Richard Bowden

Hand pose estimation from a single image has many applications. However, approaches to full 3D body pose estimation are typically trained on day-to-day activities or actions. As such, detailed hand-to-hand interactions are poorly represented, especially during motion. We see this in the failure cases of techniques such as OpenPose or MediaPipe. However, accurate hand pose estimation is crucial for many applications where the global body motion is less important than accurate hand pose estimation.
This paper addresses the problem of 3D hand pose estimation from monocular images or sequences. We present a novel end-to-end framework for 3D hand regression that employs diffusion models that have shown excellent ability to capture the distribution of data for generative purposes. Moreover, we enforce kinematic constraints to ensure realistic poses are generated by incorporating an explicit forward kinematic layer as part of the network. The proposed model provides state-of-the-art performance when lifting a 2D single-hand image to 3D. However, when sequence data is available, we add a Transformer module over a temporal window of consecutive frames to refine the results, overcoming jittering and further increasing accuracy.
The method is quantitatively and qualitatively evaluated showing state-of-the-art robustness, generalization, and accuracy on several different datasets.