

# 1 Introduction

## 1.1 Related Work

2-dimensional pose estimation can be divided into either being image-based or video-based, where the methods in the latter case use the tempoeral information of the video to perform the pose estimation.

Image-based methods were initially based on the geometry between the joints of the taget image [10, 11, 16]. Following this, were the convolutional-based methods, that used convolutional neural networks [6] to perform the pose estimation [13, 8, 2, 4]. More recent methods use transformers [12] to deliver state-of-the-art results [14, 15].

Early video-based methods used 3-dimensional convolutions to capture the temporal information between neighboring frames [9, 3]. Other methods use LSTM's [5] to capture this temporal information [7, 1]. Like in the case of image-based methods, transformers [12] have recently been introduced to the video-based methods to capture the temporal information and deliver state-of-the-art-results [17].

## 1.2 Problem Definition

## 1.3 Reading Guide