Group Spatial Attention for 3D Human Pose Estimation

Tien-Dat Tran, Ge Cao, Russo Ashraf and Kang-Hyun Jo
School of Electrical Engineering, University of Ulsan
Ulsan (44610), South Korea
Email: ttd9x1995@gmail.com, caoge9706@gmail.com, ashrafrusso@gmail.com, acejo@ulsan.ac.kr

***Abstract*—This paper introduces a novel Group Spatial Attention Module (GSAM) for enhancing 3D Human Pose Estimation (3DHPE) accuracy in complex scenes. Traditional 3DHPE approaches often struggle with occlusions and varied human poses, leading to decreased precision. GSAM addresses these challenges by leveraging group spatial attention mechanisms that dynamically focus on relevant spatial features and interactions among multiple figures within a scene. Our method incorporates a deep learning architecture that integrates GSAM with a state-of-the-art 3DHPE framework, facilitating the extraction of rich, contextual spatial information. We evaluate our approach on standard benchmarks, including Human3.6M and MPI-INF-3DHP, demonstrating significant improvements over existing methods in terms of accuracy and robustness against occlusions and pose variations. GSAM sets a new standard for 3DHPE, offering substantial advancements for applications in augmented reality, surveillance, and interactive systems.**

***Index Terms*—3D Human pose estimation, efficient attention module, transformer.**

I. Introduction

The advent of 3D Human Pose Estimation (3DHPE) has marked a pivotal advancement in computer vision, offering profound implications for various applications, including augmented reality, sports analysis, human-computer interaction, and surveillance. Despite significant progress, accurately estimating 3D human poses in complex environments remains a formidable challenge due to factors such as occlusions, the diversity of human poses, and interactions among multiple individuals.

Background and Challenges: Early attempts at 3DHPE were primarily focused on controlled environments with minimal occlusions and interactions. However, real-world applications demand robust performance in much more complex scenarios. Traditional methods often rely on single-frame analysis or simplistic spatial feature extraction techniques, which are not sufficient to handle the intricate dynamics of real-life scenes.

The Emergence of Spatial Attention Mechanisms: Recognizing the limitations of conventional approaches, recent research has turned to spatial attention mechanisms as a means to enhance feature extraction by dynamically prioritizing regions of interest within an image. These methods have shown promise in improving the accuracy of 3DHPE by enabling models to focus on relevant features while minimizing the impact of occlusions and irrelevant background information.

Introducing Group Spatial Attention Module (GSAM): Building on the foundation of spatial attention, we propose the Group Spatial Attention Module (GSAM), a novel component designed to revolutionize 3DHPE by specifically addressing the challenges posed by group interactions and occlusions in complex scenes. Unlike traditional attention mechanisms that treat figures independently, GSAM considers the spatial relationships and dependencies among multiple figures, enabling a more nuanced understanding of the scene.

Technical Overview: GSAM integrates seamlessly with existing 3DHPE frameworks, employing a deep learning architecture that leverages both global and local spatial contexts. It utilizes group-wise attention layers to dissect and analyze the spatial dynamics among individuals within a scene, enhancing the model's ability to discern occluded or closely interacting figures. This is achieved through a sophisticated algorithm that dynamically adjusts the focus of attention based on the configuration and orientation of figures concerning each other.

In summary, the main contribution of the paper is described in two-fold:

- We design and apply a new module called the group spatial attention that makes the data of 2D Keypoint can solve the occluded problem.
- We comprehensively evaluate and compare the proposed method with the original method on the Human3.6M and MPII-INF-3DHP benchmark dataset, which is the most popular dataset for keypoint.

II. Related work

2D-Human Pose Estimation Joint detection and its relationship to spatial space are the most crucial elements of human pose estimation, as shown in Fig. 2. The bottom-up method and the top-down method are the two basic approaches used for estimating human pose. Simple baseline uses joint prediction for the bottom-up technique, Deeppose [1], employing an end-to-end network with a higher parameter. Later, Newell minimizes the number of settings while keeping high accuracy by using the Stacked hourglass network [2]. All the approaches used Gaussian distributions to model local joints. An estimation of human posture was then performed using a convolution neural network. For the top-down method, first, we apply a detector for the human proposal region, and after that, we use the crop region for pose estimation. Because the top-down method uses the detector the accuracy can be better than the bottom-up. And bottom-up is an end-to-end method so the inference time can be better than the top-down.