# 2D to 3D Pose Lifting

Mohammadamin Samadi Khoshkhoo
University of Alberta
msamadik@ualberta.ca

Mohammadali Shakerdargah
University of Alberta
shakerda@ualberta.ca

Amirhosein Ghasemabadi
University of Alberta
ghasemab@ualberta.ca

## 1. Team members' contributions

In this project, we will use 2D images to construct a 3D pose of the subject. The first step of our work was to study some related papers and available public datasets. Each group member was responsible for reviewing two papers, six papers in total.

### 1.1. Mohammadamin Samadi Khoshkho

- *Exploiting Temporal Contexts with Strided Transformer for 3D Human Pose Estimation* [2]:
  This study proposes an innovative architecture for constructing a 3D pose from a long sequence of 2D joint locations. The proposed architecture, Strided Transformer, captures the dependencies of 2D pose sequences using Vanilla Transformer Encoder (VTE). To shorten the sequence, strided convolutions are substituted for the Vanilla Transformer Encoder's fully-connected layers in the feed-forward network. This architecture's performance on two benchmarks, Human3.6M and Human Eva-I, was superior to that of other state-of-the-art methods.

- *Semantic Graph Convolutional Networks for 3D Human Pose Regression* [6]:
  This study presents a novel architecture called Semantic Graph Convolutional Network that attempts to capture semantic information and applies this architecture to the regression problem of the human 3D pose. This method is created for use with single image inputs, not videos. The primary reason for employing Graph Convolutional Networks is that CNNs require grid-like inputs, whereas many tasks have irregular input structures.

### 1.2. Mohammadali Shakerdargah

- *Cascaded Deep Monocular 3D Human Pose Estimation with Evolutionary Training Data* [1]:
  This paper proposes a framework to enrich the 3D pose distribution of an initial biased training set. A novel cascaded 3D human pose estimation model is trained to achieve state-of-the-art performance for single-frame 3D human pose estimation. Many fruitful directions remain to be explored. Extension to temporal domain, multi-view setting and multi-person scenarios are three examples. In addition, instead of being fixed, the operators can also evolve during the data generation process. The method that has been used outperforms other Monocular 3D Human Pose Estimation in a lot of metrics.

- *Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image* [3]:
  This paper proposes a unified formulation for the problem of 3D human pose estimation from a single raw RGB image that reasons jointly about 2D joint estimation and 3D pose reconstruction to improve both tasks. An integrated approach has been used that fuses probabilistic knowledge of 3D human pose with a multi-stage CNN architecture and uses the knowledge of plausible 3D landmark locations to refine the search for better 2D sites which substantially outperform all other methods in terms of average error, showing a 4.7mm average improvement over our closest competitor.

### 1.3. Amirhosein Ghasemabadi

- *SRNet - Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach* [4]:
  Human rare or unseen poses in a training set are challenging for a network to predict. This paper proposed a novel Dynamical Graph Network (DG-Net), which can dynamically identify human-joint affinity, and estimate 3D pose by adaptively learning spatial/temporal joint relations from videos. The proposed Dynamical Spatial/Temporal Graph convolution discovers spa-

tial/temporal human-joint affinity for each video exemplar, depending on spatial distance/temporal movement similarity between human joints in this video. Hence, it can effectively understand which joints are spatially closer and have a consistent motion for reducing depth ambiguity and movement uncertainty when lifting 2D pose to 3D pose.

- *MixSTE - Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video* [5]:
  Some recent papers consider body joints among all frames globally to learn spatio-temporal correlation. This paper proposes MixSTE (Mixed Spatio-Temporal Encoder), which has a temporal transformer block to model each joint's temporal motion and a spatial transformer block to learn inter-joint spatial correlation. These two blocks are utilized alternately to obtain better spatio-temporal feature encoding. In addition, the network output is extended from the central frame to the entire frames of the input video, thereby improving the coherence between the input and output sequences.

## 2. Achievements so far

After reading six reputable papers on the subject of 2D to 3D pose lifting using various techniques and procedures and looking into different neural network structures, we came to the conclusion that we should use Strided Transformer [2] because it makes use of the attention mechanism. We think there is a good chance of improving the neural network's structure and algorithm to achieve a better performance.

## 3. Obstacles

We utilize a sequence of 2D images rather than just one image for a specific event, and because of this, one challenge we are facing is determining whether or not recurrent networks will be helpful for the estimation. Since we use a transformer, a major part of the process is parallel. By adding recurrent layers, however, we would be adding a lot of sequential computations, which would slow down the speed of learning. We are having some difficulty balancing the trade-off while employing recurrent layers.

## 4. Goal

Our objective is to enhance the performance of recent research on 2D to 3D pose lifting by utilizing attention mechanisms and different methods in neural networks.

## 5. Schedule

We will spend half of our time modifying the neural network architecture, looking into ways to boost performance, and creating a presentation. After that, we will spend our time writing and editing the final report.

## References

[1] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1

[2] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation, 2021. 1, 2

[3] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017. 1

[4] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach, 2020. 1

[5] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video, 2022. 2

[6] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. 1