Lifting 2D-3D Human Pose Estimation System with Attention and Transformer Techniques in Industrial Environment

***Abstract*— Effectively reducing a redundant 2D pose sequence from a weak pose detector to create a representative 3D pose is still an unresolved challenge. To address this, the proposed method introduces an efficient system incorporating the attention mechanism. Inside the system, two main networks are deployed: the first is the 2D pose detector, and the second is the 3D pose estimator. Additionally, the 2D pose estimator utilizes an attention module to achieve precise 2D joint detection. To enhance the accuracy of the 2D pose, a new attention module is implemented after each block. Concerning the 3D pose network, it also improves the Transformer-based architecture through the advanced attention mechanism. Specifically, a new Transformer Encoder, applies spatial and temporal attention, is to capture long-range dependencies in 2D pose sequences. The proposed architecture has demonstrated state-of-the-art performance on two benchmark 3D human pose estimation datasets, Human3.6M and MPI-INF-3DHP. Notably, this research enhances performance by 0.9% and 0.3%, respectively, when compared to the closest counterpart, PoseFormer. In terms of 2D pose, the proposed system also beats current methods on the COCO 2017 Microsoft Dataset. Link demo: demo vision**

***Index Terms*— 3D modeling, Pose estimation, Deep learning, Video surveillance.**

I. Introduction

A. Research Background

THREE Dimension human pose estimation is a crucial topic in computer vision. This approach involves determining the three-dimensional locations of a human body joint from a two-dimensional image or set of photos. Many applications can be used for human pose estimation such as object recognition [1], [2], human-computer-interaction [3], activity recognition [4], [5] or robotic system [6], [7].

1) 2D Human Pose Estimation: Due to the wide range of applications mentioned in the introduction part, most of these methods fall into two categories: top-down methods and bottom-up methods. In top-down methods, a human detector is generally used to detect all the people in the image first, and then single-person pose estimation is conducted for each detected subject separately. Bottom-up methods [27] have also attracted a lot of attention recently due to its efficiency. Besides bottom-up methods, The single-person pose estimation methods that are commonly used in top-down methods include HRNet [18], and HRFormer [8], etc. In top-down methods, most methods first detect the human proposal and then detect the join inside. Differently, This paper proposed a new bottom-up method that optimizes the heatmap prediction via applying the attention mechanism between the characteristic functions of the predicted and GT heatmaps.

2) 3D Human Pose Network: Existing single-view 3D pose estimation methods can be divided into two mainstream types: one-stage approaches and two-stage ones. One-stage approaches directly infer 3D poses from input images without intermediate 2D pose representations [19], [29], while two-stage ones first obtain 2D keypoints from pretrained 2D pose detections and then feed them into a 2D-to 3D lifting network to estimate 3D poses. Benefiting from the excellent performance of 2D human pose estimation, this 2D-to-3D pose lifting method can efficiently and accurately regress 3D poses using detected 2D keypoints. Despite the promising results achieved by using temporal correlations from fully convolutional [4], [26] or graph-based [2] architectures, these methods are less efficient in capturing global-context information across frames.

Recently, vision transformers advanced all the visual recognition tasks [14]. Following PoseFormer [21], transformer has been used to lift 2D poses to the corresponding 3D poses. To eliminate the redundancy in the sequence with temporal information, Li et al. [12] proposed a strided transformer network. spatial-temporal transformer is used for 3D HPE tasks. Using transformers in HPE showed significant improvement overall. However, pre-training on a large dataset is required to learn more representative and effective representations for the sequence HPE data. The proposed method is different from the previous methods in leveraging the cross-interaction between the joints of the body parts in the spatial and temporal domains.

B. Problem Statement and Technical Challenges

For 2D pose Estimator, deep convolution neural networks have achieved outstanding performance. Before raising the resolution, most existing techniques route the input through a network and, after that apply the 3D human pose estimation on the 2D result, which is shown in Fig.1. The 3D network takes the series of 2D points as the input and is typically made up of high-to-low resolution sub-networks connected in series. Hourglass [11], for example, uses a symmetric low-to-high technique to recover high resolution. Simple Baseline [27] uses a few transposed convolution layers to build high-resolution representations. Hence, the accuracy of the 2D key point and lifting is to 3D model still a big problem.

Recent advancements in 3D human posture encoding have been achieved through deep neural networks [17], [22]. However, these networks face numerous obstacles. Firstly, how can the accuracy of different network types, such as real-time networks or correctness-measuring networks, be improved? Secondly, it is common to check the accuracy of a network while utilizing different 2D pose results. Finally, the current network must increase accuracy while maintaining the speed or make it as fast as possible. The proposed study investigates a unique network and evaluates it with speed and accuracy. Hence, this experiment diverges from PoseFormer [21] by applying the new attention mechanism which calls spatial-temporal attention.

The proposed technique was employed to develop a 3D pose network, demonstrating a noteworthy enhancement in Mean Per Joint Position Error (MPJPE). The proposed network, inspired by PoseFormer [21], aims to enhance the attention mechanism for the 2D keypoint using the attention module. By employing a new spatial and temporal attention module inside the baseline transformer, the network keeps the MPJPE higher while minimizing the implementation cost. In addition, the number of parameters was reduced, which resulted in a faster network. In gaining a deeper understanding of the attention within the transformer, the proposed method reduces the Average MPJPE error by 1.0 points.

C. Attention in Human Pose Estimation Review

The attention mechanism has been widely adopted in natural language processing (NLP) tasks, to achieve state-of-the-art performance in the machine translation task [5] and language understanding task [21] Recently, attention-aware features are also found to work very well in computer vision tasks. For example, [11] proposed a powerful attention module that combines an attention branch with an hourglass block, which was stacked multiple times to form a deep convolutional neural network for image classification. Based on the self-attention mechanism, the network proposed in [13] captured rich contextual dependencies for the scene segmentation task. [16],[11] incorporated the attention mechanism with different convolutional neural networks for human pose estimation. One popular attention mechanism used in human pose estimation is self-attention, also known as transformer-based attention. Self-attention allows the model to attend to different parts of the input and capture long-range dependencies. By applying self-attention to pose estimation, models can dynamically weight the importance of different joints or body parts based on their relationships. Additionally, spatial attention can be employed to highlight relevant spatial regions in an image. Spatial attention mechanisms, such as spatial transformer networks or spatial attention modules, enable models to attend to informative image regions that are crucial for pose estimation.

D. Contribution of The Paper

Many papers have been researched on 2D and 3D human pose estimation over the past few years. However, less work has been deeply studied on attention mechanisms for both 2D and 3D networks. This article proposes a new attention mechanism for the whole network, which significantly improves the accuracy of the final 2D and 3D prediction results. In summary, the main contribution of the paper is described in three-fold:

1) This paper proposed and applied the attention mechanism not only for the 2D pose detector but also for the 3D estimator, making the network emphasize more information about the occluded problem.
+ In 2D Pose Network, an attention module that applies depth-wise convolution.
+ Proposed new spatial-temporal attention for The 3D Network, which increases the accuracy of the 3D prediction.
2) Approach a new two-stage network that makes the input image into 3D human pose prediction. Also applying some small techniques to the two-stage network to improve both 2D and 3D result
3) Without bells and whistles, our proposed method outperforms the original method on the benchmark dataset.
+ For 2D, comprehensively compare other methods in the Microsoft COCO 2017 benchmark
+ Have competitive results for Human 3.6M and MPI-INF-3DHP dataset for 3D Network.

II. Methodology

A. 2D Pose Estimator

1) Backbone network: Proposed system utilized a benchmark comprised of HighResolutionNet-W32 and HighResolutionNet-W48 [18], as depicted in Fig. 1A for a complete network. Each HighResolutionNet is divided into four stages four subnetworks that contain skip connections and residual blocks. The default data image is reduced in dimension to 256 × 192 (HighResolutionNet-W32, HighResolutionNet-W48), the extracted feature traverses each stage, with the starting dimension of $H \times W$ is reduced twice for each stage. After the data travels down the till the end of the backbone, the extract map size is reduced to $\frac{W}{16} \times \frac{H}{16}$ with 256 channels at final layer. Therefore, the backbone architecture will only employ the first subnet, whose dimension remains $W \times H$ until the conclusion of the regression. Additionally, the channels' dimensions were increased 2 times at each level. The tensor channel increased from 32 at the first stage to 256 at the end. The baseline architecture's role is to collect valuable information from extract tensor and provide it to the training process, which predicts human joints via cross-entropy loss.

2) Attention Module: In the proposed 2D Pose Estimator, the Attention Mechanism was applied after each stage of the first sub-network and second sub-network as same as In Fig.1A Only 4 attention modules were utilized to balance the computational cost and accurate. In Fig. 2, the attention module consists of two main parts. After block one in the backbone network, the feature information was first transferred to the channel attention module (CAM). The tensor information in CAM utilizes a GAP layer (Global average Pooling) to decrease the tensors from $W \times H \times C$ to $1 \times 1 \times C$. Then, It traverse through the convolution layer, which made the important feature into $1 \times 1 \times \frac{C}{r}$, where the shrinking ratio is r which set at default to 16. The weight inside network was

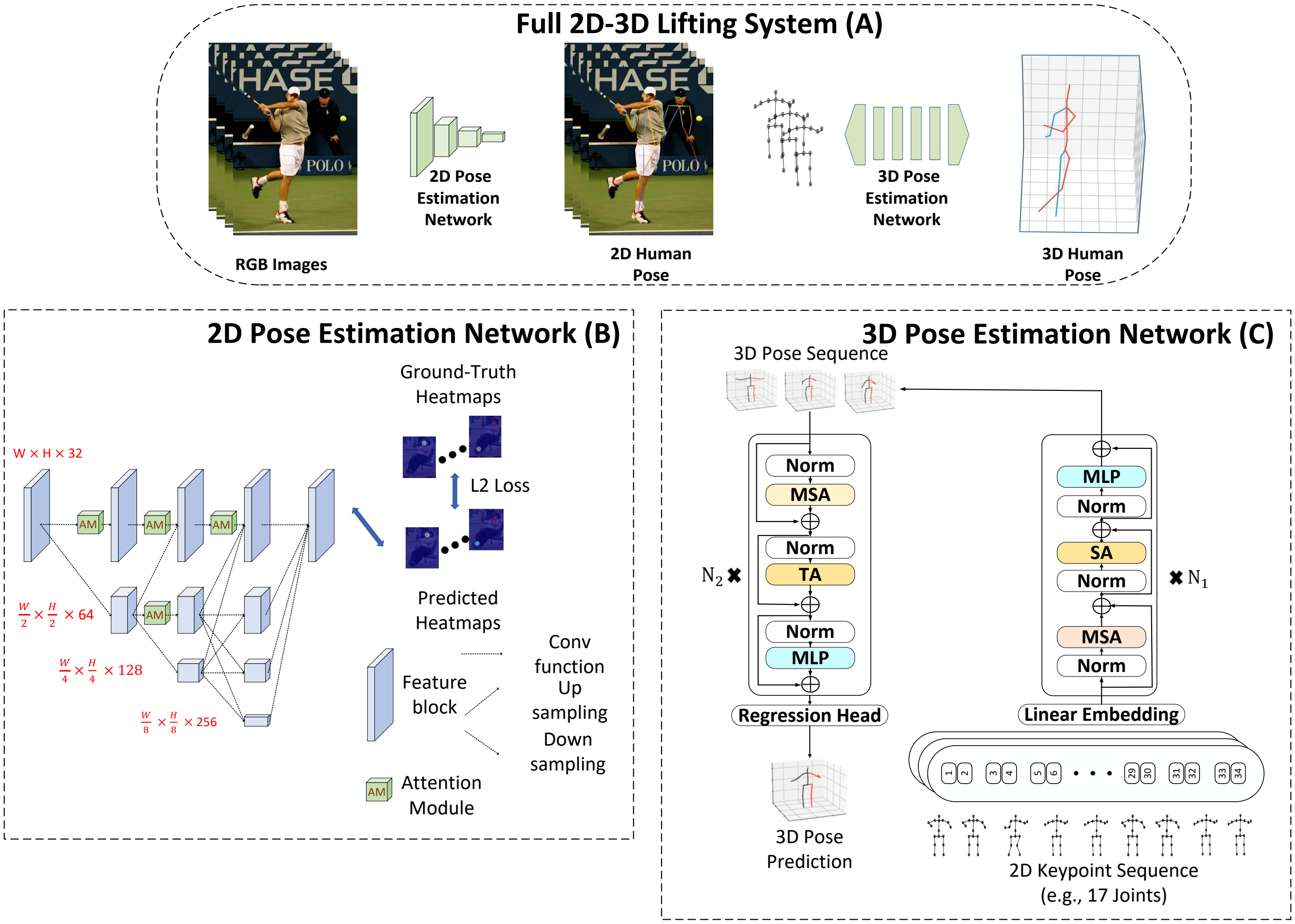

Fig. 1. Illustrate from 2D to 3D human pose estimation. The proposed method separated the system (A) into two networks, B is for 2D Pose Estimator and C is 3D Pose Network

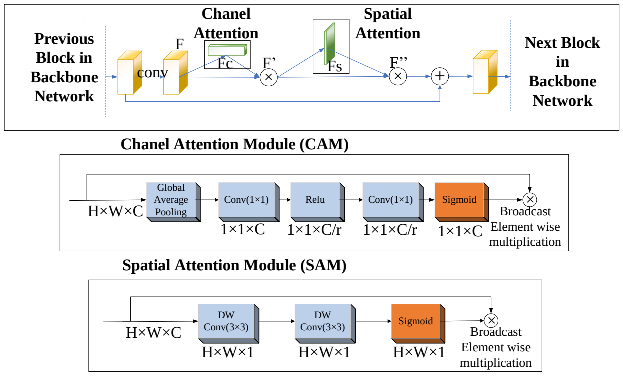
Previous Block in Backbone Network
Chanel Attention
Spatial Attention
Next Block in Backbone Network
F
conv
Fc
F'
Fs
F''
Chanel Attention Module (CAM)
H×W×C
Global Average Pooling
Conv(1×1)
Relu
Sigmoid
1×1×C
1×1×C/r
Broadcast Element wise multiplication
Spatial Attention Module (SAM)
DW Conv(3×3)
H×W×1

Fig. 2. **Top:** Full Attention Module architecture. **Middle:** Chanel Attention Module **Bottom:** Spatial Attention Module

triggered by the Channel mechanism utilizing function ReLU for the activation. Finally, the proposed CAM utilize a 1×1 convolution layer to made the size of channel to $1 \times 1 \times C$ and apply sigmoid to normalize the weight in final tensor. The data in CAM were then mixed by utilize a multiplication of element-wise.

After going through the CAM, the tensor will be sent into the SAM. The tensors in the Spatial module change the channel's by applying average pooling so the tensor from $W \times H \times C$ to $W \times H \times 1$. The last step in SAM is supplied to the CAM depicted in Figure 2 after pooling, and convolution layers with 3×3 kernel size were used twice to extract the geographic data for the network.

B. 3D Pose Estimation Network

1) Baseline network: In this work, it adopt a Transformer-based architecture which in Fig. 5 since it performs well in long-range dependency modeling. Then first give a brief description of the basic components in the Transformer [13], including a multi-head self-attention(MSA) and a multi-layer perceptron (MLP). MSA. In the MSA, the inputs $x \in \mathbb{R}^{n \times d}$ are linearly mapped to queries $Q \in \mathbb{R}^{n \times d}$, keys $K \in \mathbb{R}^{n \times d}$, and values $V \in \mathbb{R}^{n \times d}$, where n is the sequence length, and d is the dimension. The scaled dot-product attention can be computed by:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_m}})V, \qquad (1)$$

MSA splits the queries, keys, and values for h times as well as performs the attention in parallel. Then, the outputs of the attention heads are concatenated. The MLP consists of two linear layers, which are used for non-linearity and feature transformation:

$$MLP(x) = \alpha(xW_1 + a_1)W_2 + a_2, \tag{2}$$

where α denotes the GELU activation function, $W_1 \in \mathbb{R}^{d\times d_m}$ and $W_2 \in \mathbb{R}^{d_m\times d}$ are the weights of the two linear layers respectively, and $a_1 \in \mathbb{R}^{d_m}$ and $a_2 \in \mathbb{R}^{d}$ are the bias terms.

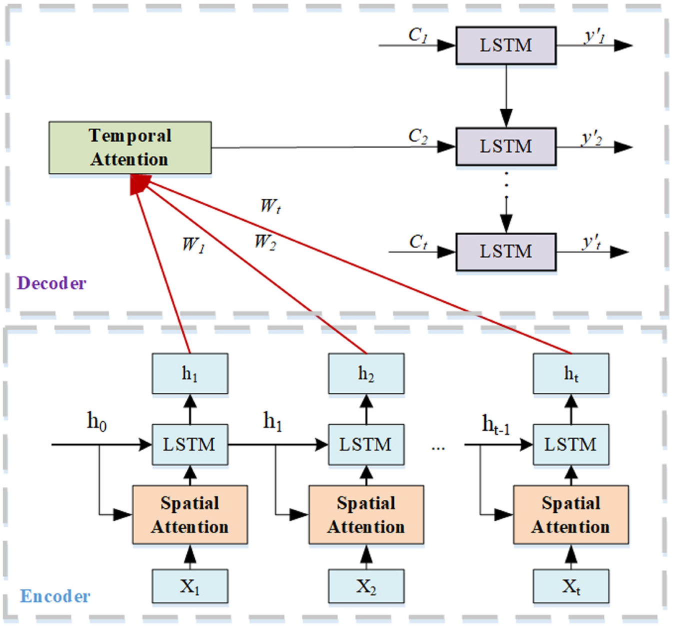

Fig. 3. Spatial and Temporal Attention inside the 3D Network

2) Spatial Attention (SA module): This module is inserted between the MSA layer and MLP for each block. The Spatial attention module consists of two depth-wise convolutions with kernel size 5, group normalization and non-linearity GELU. Also, the residual connection is added to the output of the module to avoid overfitting. The following operations on output of the patch embedding step P_0 can be described:

$$P = CONV(Norm(GELU(CONV(P)))) + P, \tag{3}$$

where $GELU$ refers to the non-linear layer, $CONV$ is the standard convolution layer with kernel 5 and Norm indicates th normalization used in [28]. Since the focus of the SA module is on the interaction between the joints, the output of the MSA part in Eq.2 has been transposed. That is, it becomes $P_0 \in R$ D×P . The spatial encoders for a transformer layer l cn then be represented by the following list of operations:

$$MLP(x_0) = \beta(xW_1 + a_1)W_2 + a_2, \tag{4}$$

where β denotes the P function in Eq.3

3) Temporal Attention (TA module): Same with SA module, The TA module learns pairwise feature correlations using the outer product. Each element of the correlation matrix $C_{ij} = \sum_F P_i P_j$ is a dot product of the corresponding embedded features of frames i and j and then it is sum-pooled, where $P_i \in R^{J\times D}$ is the input feature of frame i. More precisely, the input is transformed by combining the positional information with the frames where $P \in R^{F\times J\times D}$ and then using convolutions this paper extract K, Q, and V such that:

$$K = PW_k, Q = PW_q, V = PW_v \tag{5}$$

4) Regression Head: In the regression head, a linear transformation layer is applied on the output Z_{L3} to perform regression to produce pose sequence $\widetilde{X} \in \mathbb{R}^{N\times J\times 3}$. Finally, the 3D pose of center frames $\widehat{X} \in \mathbb{R}^{J\times 3}$ is selected from $\widetilde{X}$ as our final prediction

C. Loss Function

1) 2D Pose Estimator Loss: Heat maps are utilized in the proposed work to demonstrate body keypoint locations in the loss function. At the beginning, we set the ground-truth point by $m = \{m_n\} N = 1^N$, where $X_n = (x_n, y_n)$ is the geographical information of the n^{th} body keypoint for every image. The principles of Ground-truth heat map H_n is then build up by utilized the Gaussian distribution and the mean a_n with variance $\sum$ as illustrated in the next equation.

$$H_n(p) \sim N(a_n, \sigma), \tag{6}$$

where $\mathbf{p} \in \mathbb{R}^2$ illustrate the coordinate, and σ is automatically decided as an identity matrix $\mathbf{I}$. The final layer of the proposed architecture generated J heat maps, *i.e.,* $\hat{S} = \left\{\hat{S}n\right\} n = 1^N$ for K body joints. Mean square error for loss function is defined, which is summarized as follows:

$$L = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left\|S_n - \hat{S}_n\right\|^2, \tag{7}$$

2) 3D Loss: The total Loss for 3D network compose in

$$\mathcal{L} = \mathcal{L}_g + \lambda_t\mathcal{L}_t + \lambda_m\mathcal{L}_m, \tag{8}$$

where $\mathbf{p} \in \mathbb{R}^2$ demonstrate the coordinate, and $\sum$ is experimentally decided as an identity matrix $\mathbf{I}$. The last layer of the neural architecture forecast J heat maps, *i.e.,* $\hat{S} = \left\{\hat{S}j\right\} j = 1^J$ for J body joints. A L2 loss function is defined by the mean of MPJPE, which is calculated as follows: The entire model is trained an end-to-end manner with a Mean Squared Error (MSE) loss, which is used for both 2D pose detector network and 3D pose network. For the 3D pose network, the Loss function

$$\mathcal{L} = \sum_{m=1}^{M}\sum_{j=1}^{J}\left\|S_j^n - \hat{S}_j^n\right\|_2, \tag{9}$$

M denotes the number of selected in the training process. Using 3D pose data from the last layer or backbone architecture, the trained network generated predict 3D joint maps using ground-truth 3D pose.

III. Experiment

A. Datasets and Evaluation Protocols

For the 2D human pose estimator, Microsoft COCO 2017 [3] was used for training and testing in the whole process.

TABLE I
COMPARISON RESULT ON COCO 2017 VALIDATION SET.

Methodology	Backbone	#Parameters	Image dimension	AP	AR	AP^{50}	AP^{75}	AP^{L}	AP^{M}
Fine-tune Attention [26]	ResNet-50	31.2M	256×192	71.4	76.3	91.6	78.6	75.7	68.2
Fine-tune Attention [26]	ResNet-101	50.2M	256×192	72.3	77.1	92.0	79.4	77.1	68.3
High-Resolution Net [18]	HighResolutionNet-W32	28.5M	256×192	74.4	79.8	90.5	81.9	81.0	70.8
High-Resolution Net [18]	HighResolutionNet-W48	63.6M	256×192	75.1	80.4	90.6	82.2	81.8	71.5
Zhang at al. [17]	HRNet-W32	29.2M	256×192	74.8	77.6	92.5	81.6	79.3	72.0
Zhang at al. [17]	Hourglass-8	25.8M	256×192	75.1	80.4	90.6	82.6	81.9	71.6
MogaNet-T [25]	MogaNet	8.1M	256×192	73.2	90.1	81.0	78.8	-	-
MogaNet-S [25]	MogaNet	29M	256×192	74.9	90.7	82.8	80.1	-	-
PPE-Net [24]	ResNeXt-101	-	256×192	75.7	-	90.3	76.3	80.7	79.5
Our	HighResolutionNet-W32	31.2M	256×192	75.7	80.6	90.6	82.1	82.4	71.3
Our	HighResolutionNet-W48	66.9M	256×192	76.1	80.9	90.7	82.7	82.9	71.9

1) Microsoft COCO 2017: was utilized through the training and testing process. This dataset is a challenging dataset for joint detection which comprises around 250K human labeled in 200K images, each human pose have 17 keypoint labels. The proposed research applies Object Keypoint Similarity (OKS) for Microsoft COCO2017 dataset with $OKS = \frac{\sum_i exp(-{d_i}^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)}$ In the above function, The Euclidean distance between the groundtruth joint and the predicted joint is d_i , The target's visibility flag is v_i, The object scale is s, and k_i is one of seventeen joints in Microsoft COCO 2017 Benchmark. Hence, The standard average accuracy and recall value are then computed.

About the 3D human pose, this approach evaluate proposed model on two general datasets: Human3.6M [9], MPI-INF-3DHP [20] and Industrial dataset individually.

2) Human3.6M: is the most commonly used indoor dataset for the 3D human pose estimation tasks. Following the same policy of the base method [14], the 3D human pose in Human3.6M is adopted as a 17-joint skeleton, and the subjects S1, S5, S6, S7, S8 from the dataset are applied during training, the subjects S9 and S11 are used for testing. The two commonly used evaluation metrics (MPJPE and P-MPJPE) are involved in this dataset. In addition, mean per-joint velocity error (MPJVE) is applied to measure the smoothness of the prediction sequence.

3) MPI-INF-3DHP: is a recently proposed large-scale dataset, which consists of three scenes, i.e., green screen, non-green screen, and outdoor. By using 14 cameras, the dataset records 8 actors performing 8 activities for the training set and 7 activities for evaluation. Following the works [21], proposed network adopt the MPJPE (P1), percentage of correct keypoints (PCK) with 150mm, and area under the curve (AUC) results as the evaluation metrics.

4) ISLAB Industrial dataset: total 4 videos which recorded people working inside the industrial environment. There are 9980 frames with 5 people inside the video. The dataset not provide the grouthtruth so the network utilizes this dataset for testing.

B. Implementation Details

The proposed model is implemented with Pytorch that use 2D keypoints from HRNet detector [18], CPN Detector or 2D ground truth to analyze the performance. In this paper, the 2D pose detector was implemented based on AlphaPose [15] codebase while the 3D pose estimator followed the PoseFormer codebade [21]. Although the proposed model can easily adapt to any length of the input sequence, to be fair, we select some specific sequence lengths T for three datasets to compare our method with other methods which must have a certain 2D input length: Human3.6M (T=81, 243), MPI-INF-3DHP (T=1, 27). Analysis about the frame length setting is discussed in the ablation study Section III.E.3. The batch size, dropout rate, and activation function for datasets are set to 1024, 0.1, and GELU. This proposed architecture utilizes the stride data sample strategy with interval is as same as the input length to make there no overlapping frames between sequences(more details in the supplementary material). All experiments are implemented on the PyTorch framework with two NVIDIA Geforce GTX 2080 Ti. The network is trained using Adam optimizer [10]. The learning rate is 0.001 with a shrink factor is 0.95 after 2 epochs. The learning rate is also this paper's contribution, which is explained in Section III.E.3.

C. Comparison with the SOTA 2D Pose Methods

1) Result for COCO2017 dataset: The proposed result in Table I was estimated on the COCO validation dataset. In all instances, the accuracy in the proposed technique is larger than the Benchmark High-Resolution Network of 1.3 and 1.0 AP in backbone HRNet-32 and HRNet-W48 respectively. In addition, the average recall (AR) for HRNet-W32 is 0.5 points higher and 0.4 points higher for HRNet-W48. Overall, the experiment outcomes improved modestly in both AP and AR, showing that attention mechanisms affect the result.

D. Comparison with the SOTA 3D Pose Methods

1) Result for Human3.6M dataset: For the 2D-to-3D pose lifting task, the accuracy of the 2D detections directly. In order to guarantee fair comparisons, the input is taken from CPN in the form of 2D keypoints for training and testing. Table I shows the comparison of the SOTA methods with the proposed method (81 frames). In Table II, the proposed method achieves the state-of-the-art on Human3.6 on all the metrics and it outperforms the state-of-the art (Chen at al) with a considerable margin of 0.9%, 1.3% for Protocols 1 and 2, respectively. It is worth noting that the across-joint modules in the spatial and temporal cases are crucial to

TABLE II

QUANTITATIVE COMPARISONS WITH STATE-OF-THE-ART ON HUMAN3.6M DATASET USING CPN DETECTOR UNDER PROTOCOL #1 AND PROTOCOL #2 FOR FULLY-SUPERVISED METHODS. BOLD NUMBER IS THE BEST PERFORMANCE IN EACH CASE. † DENOTE OUR 3D NETWORK APPLY OUR 2D NETWORK

Protocol # 1 - CPN	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Punch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [16]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang *et al.* [22]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Li *et al.* [23]	47.0	**47.1**	49.3	50.5	**53.9**	**58.5**	48.8	45.5	**55.2**	68.6	50.8	**47.5**	53.6	42.3	45.6	50.9
Zhen [18]	45.4	49.2	45.7	49.4	50.4	58.2	**47.9**	46.0	57.5	63.0	49.7	46.6	52.2	**38.9**	40.8	49.4
Xu *et al.* [19]	45.2	49.9	47.5	50.9	54,9	66.1	**48.5**	46.3	59.7	71.5	51.4	48.6	53.9	**39.9**	44.1	51.9
Yang *et al.* [21]	41.5	44.8	39.8	42.5	46.5	**51.6**	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	**44.3**
Our	**45.0**	48.3	**46.6**	**49.8**	46.6	59.0	48.7	**41.9**	57.7	**60.2**	**45.1**	48.2	**45.8**	41.0	45.1	**43.1**
Protocol # 2 - CPN	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Punch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Fang *et al.* [22]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Pavlakos *et al.* [29]	34.7	39.8	41.8	**38.6**	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Yang *et al.* [30]	**26.9**	**30.9**	**36.3**	39.9	43.9	47.4	**28.8**	**29.4**	**36.9**	58.4	41.5	**30.5**	**29.5**	42.5	**32.2**	**37.7**
Yang *et al.* [21]	30.0	33.6	29.9	31.0	30.2	35.4	37.4	34.5	46.9	**50.1**	40.5	36.1	41.0	29.6	33.2	39.0
Li *et al.* [23]	34.5	34.9	37.6	39.6	**38.8**	45.9	34.8	33.0	40.8	51.6	38.0	35.7	40.2	30.2	34.8	38.0
Our	34.1	36.0	36.4	39.9	39.4	**45.0**	35.9	32.8	43.1	52.1	**37.3**	36.6	39.7	30.2	35.8	38.3

TABLE III

QUANTITATIVE COMPARISONS WITH STATE-OF-THE-ART ON HUMAN3.6M DATASET USING GROUNDTRUTH AS 2D KEYPOINT UNDER PROTOCOL #1 WITH 2D GROUND-TRUTH INPUT. BOLD NUMBER IS THE BEST PERFORMANCE IN EACH CASE

Protocol # 1 - GrouthTruth	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Punch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [16]	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Fang *et al.* [22]	32.1	36.6	34.3	37.8	44.5	49.9	40.9	36.2	44.1	45.6	35.3	35.9	30.3	37.6	35.6	38.4
Li *et al.* [23] †	32.9	38.7	32.9	37.0	37.3	44.8	38.8	36.1	41.2	45.6	36.8	37.7	37.7	29.5	31.6	37.2
Zhen [18]	45.4	49.2	45.7	49.4	50.4	58.2	47.9	31.7	38.5	45.5	35.4	36.6	36.2	28.9	30.8	35,8
Xu *et al.* [19]	35.8	38.1	47.5	31.4	39.6	35.8	45.5	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
Xue *et al.* [3]	35.0	37.2	46.6	30.8	38.7	35.1	44.3	34.9	40.1	41.0	32.1	33.6	32.5	26.0	26.1	33.3
Chen *et al.* [28]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32.3
Yang *et al.* [21]	34.8	32.1	29.8	31.5	36.9	35.6	30.5	30.5	38.9	40.5	32.5	31.0	29.9	22.5	24.6	32.0
Our	**27.9**	**29.9**	**26.6**	**27.8**	**28.6**	**32.8**	**31.1**	**26.7**	**36.5**	**35.5**	**30.0**	**29.8**	**27.5**	**19.6**	**19.7**	**31.0**

infer the body-joints dependencies. Comparing the proposed method with PoseFormer (with no pre-training used) shows the significance of the across-joint correlation modules. Our method outperforms with a large margin of 2% the SOTA. In terms of accuracy, it achieve 1% better than the second best accuracy. Additionally, the proposed method achieves the best performance amongst all the compared methods in protocol 2 in Table II (bottom). In some selected difficult poses such as walk together, walk, smoke, where the poses change very quickly, the proposed method showed a significant improvement ranging from 1.1% to 2.5% over the baseline. This highlights the ability of our method to encode the long-range interactions between the body-joints. Considering the pre-trained baseline, the proposed method achieves better performance for all the actions. These results show the importance of plugging the Spatial-temporal attention modules in the transformers.

Further experiments on Human3.6 using ground-truth 2D poses as input have also been performed. This shows the power of the proposed method where there is no noise in the input as in the previous case. Table III shows the comparisons of our method and the baselines. Overall, the proposed method achieved the best performance amongst the baselines. It achieved 28.3% MPJPE, whereas the second-best approach achieved 31.0 with gain of 3%. The proposed method outperforms the baselines in all the actions with a considerable improvement range from 2.4% as the minimum difference and 4.8% for the largest.

2) Result for MPII-INF-3DHP dataset: The approach further compares the proposed methods to previous ones on MPP-INF-3DHP using 9 frames. This is important because it illustrates the ability of the proposed method to train with fewer training samples in outdoor settings. As Table IV shows, this paper obtains the best performance amongst the compared ones w.r.t. the metrics.

TABLE IV

PERFORMANCE COMPARISION IN TERMS OF PCK, AUC AND P1 WITH THE STATE-OF-THE-ART METHODS ON MPI-INF-3DHP

Method	PCK ↑	AUC ↑	MPJPE ↓
Pavllo *et al.* [29] (f=81)	86.0	51.9	84.0
Lin *et al.* [13] (f=25)	83.6	51.4	79.8
Li *et al.* [23]	81.2	46.1	99.7
Chen *et al.* [28]	87.9	54.0	78.8
Yang *et al* [21] (f=9)	88.6	56.4	75.5
Our (f=9)	89.1	57.5	76.3

3) Result for ISLAB Industrial dataset: Fig.5 shows the 3D Human Pose Testing results on the ISLAB industrial dataset. The proposed utilize the result from the proposed 2D detector.

E. Ablation Study

1) Effect of attention in 2D Detector and 3D Estimator: In Table V, To evaluate the impact and performance of the 2D for the whole 3D model, The proposed network evaluates and investigates the result in the Human3.6M dataset. The result shows that applying the attention module in the 2D pose estimator makes the 2D input accurate and then helps the final 3D result. Fig.4 shows the impact of the attention mechanism when the arm in the picture is straight compared to the baseline HRNet looks folding the arms while in the testing image, the person is straight his arm.

TABLE V

COMPARISION RESULT FOR APPLYING THE ATTENTION MODULE IN HRNET WITH OTHER DETECTORS

Detector	Protocol #1	Protocol #2	MPJVE
CPN	47.6	37.4	3.20
Detectron2 [30]	45.7	37	3.02
Hourglass [11]	52.3	41.2	4.11
HRNet-W32 [18]	45.1	36.3	2.91
HRNet-W32+AM (our)	43.6	35.1	2.77
GroundTruth	28.6	24.5	0.98

Table VI is a comparison of different module in a proposed system, focusing on the presence or absence of specific modules and their impact on the Mean Per Joint Position Error (MPJPE). The modules include 2D Attention, 3D SAM (Spatial Attention Module), and 3D TAM (Temporal Attention Module). Each row in the table corresponds to a specific configuration, indicating the presence or absence of these modules. The MPJPE values for each configuration serve as a quantitative measure of the accuracy of joint position predictions. Notably, the proposed method exhibits improved performance when incorporating all three modules simultaneously, achieving the lowest MPJPE at 42.2, which decreases by 3.2% in accuracy comparison to the baseline.

TABLE VI

COMPARISION RESULT OF EACH MODULE IN THE PROPOSED SYSTEM

Method	2D Attention	3D SAM	3D TAM	MPJPE
PoseFormer				44.3
Our	✓			43.6
Our		✓		43.7
Our			✓	43.8
Our		✓	✓	43.3
Our	✓	✓	✓	42.2

2) Position of Attention Module in 2D Detector and 3D Estimator: Table VII investigates the result when applying different AM in each subnetwork and each stage in HRNet. In conclusion, the result when applied in the attention module in all stages (9 Attention modules got added) got the best result however it also got the highest number of parameters in the computational cost. Besides, Table VI also shows that AM had the most effect in the first sub and stage than in the remaining. Hence, this paper only applies the module for the first subnetwork and stage (only 4 was added) to not only balance the computational cost but also keep the high accuracy.

Table VIII showcases the influence of different positions of the Spatial Attention Module (SAM) and Temporal Attention Module (TAM) on Mean Per Joint Position Error (MPJPE). For SAM, positioning it after Multi-Head Self-Attention (MSA) or after Multi-Layer Perceptron (MLP) yields lower MPJPE (44.1 and 44.9) compared to before MSA (45.2). Similarly, for TAM, placing it after MSA results in the lowest MPJPE (44.9), while before MSA and after MLP have slightly higher errors (45.0 and 46.2, respectively). This highlights the importance of the relative positioning of attention modules in achieving optimal accuracy in joint position predictions. Hence, this paper decide to put SAM and TAM between the MSA and MLP.

TABLE VII

THE RESULT WHEN UTILIZING THE ATTENTION MECHANISM FOR EACH SUB-NETWORK AND EACH STAGE OF HIGHRESOLUTION NETWORK

Backbone	Sub-Net	AP	#Param
HighResolutionNet-W32	-	74.4	28.5M
HighResolutionNet-W32	1	75.4	31.1M
HighResolutionNet-W32	2+1	75.9	33.8M
HighResolutionNet-W32	3+2+1	76.3	35.5M
HighResolutionNet-W32	4+3+2+1	76.4	36.4M
Backbone	Stage	#Param	AP
HighResolutionNet-W32	1	75.5	30.2M
HighResolutionNet-W32	2+1	76.0	32.9M
HighResolutionNet-W32	3+2+1	76.4	36.4M
HighResolutionNet-W32	Sub-1 + Stage-1	75.7	31.9M

TABLE VIII

THE RESULT WHEN APPLYING DIFFERENT POSITIONS OF SAM AND TAM

Module	Before MSA	After MSA	After MLP	MPJPE
SAM	✓			45.2
SAM		✓		44.1
SAM			✓	44.9
TAM	✓			45.0
TAM		✓		44.9
TAM			✓	46.2

3) Effect of modifying the setting in 3D network: Table IX presents a comparative evaluation of different backbone architectures for human pose estimation under varying stride frame configurations. Three methods, Pavllo et al.'s approach [29], PoseFormer by PoseFormer et al. [21], and a proposed method are analyzed. For Pavllo et al.'s method, adjusting the stride frame from the default 243 to 81 leads to a slight reduction in the number of parameters from 12.75M to 12.70M, with a marginal increase in the Mean Per Joint Position Error (MPJPE) from 47.5 mm to 47.9 mm. PoseFormer demonstrates improved accuracy with reduced MPJPE values when the stride frame is decreased from 81 to 27, resulting in MPJPE values of 44.3 mm and 44.6 mm, respectively. The proposed method (”Our”) consistently outperforms the other methods, achieving lower MPJPE values as the stride frame decreases from 81 to 27 to 9, while maintaining a relatively stable parameter count of around 9.86M. This suggests that the proposed method is effective in producing accurate pose estimations with different stride frame configurations.

TABLE IX

THE RESULT FOR APPLYING DIFFERENT LEVELS OF FRAME. THE DEFAULT SETTING FOR LEARNING RATE IS 0.25

Method	Stride Frame	#Param (M)	MPJPE (*mm*)
SimplePose *et al.* [29]	243 (default)	12.75M	47.5
SimplePose *et al.* [29]	81	12.70M	47.9
PoseFormer *et al.* [21]	81 (default)	9.59M	44.3
PoseFormer *et al.* [21]	27	9.60M	44.6
Our	9	9.85M	44.3
Our	27	9.86M	43.6
Our	81	9.86M	43.3

TABLE X
THE COMPARISON RESULT FOR APPLYING DIFFERENT LEARNING RATES FOR 3D MODEL. THE DEFAULT FRAME WAS SET AT 81 FOR ALL OF THE EXPERIMENT

Method	Stride Frame	#Param (M)	MPJPE (*mm*)
SimplePose *et al.* [29]	0.25 (default)	12.70M	47.9
SimplePose *et al.* [29]	0.1	12.70M	47.5
PoseFormer *et al.* [21]	0.25 (default)	9.60M	44.3
PoseFormer *et al.* [21]	0.1	9.60M	44.6
Our	0.25	9.86M	43.3
Our	0.2	9.86M	43.3
Our	0.1	9.86M	43.1
Our	0.05	9.86M	43.4

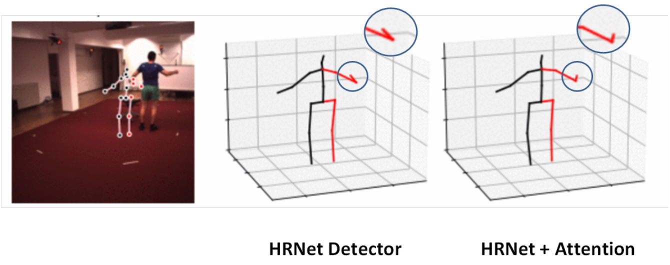

Fig. 4. 3D human pose estimation result come from 2D skeleton based on detector and detector with attention mechanism

Table X shows the result when changing the learning rate setting. While other papers set the learning rate as 0.25 and do not consider this. This paper found based on the gradient descent, 0.1 in learning rate is truly a perfect match for 3D model. Only simple changing with our increase the computational cost but significantly improve the accuracy which decreases almost 1% of the error. The side effect of changing the learning rate is only making training time increase from 20 hours to 22 hours.

IV. Conclusion

This research explores the impact of attention mechanisms not only on the 2D Pose Detector but also on the 3D Pose Estimator, particularly in the context of constructing a full system from input to 3D result for the Industrial Environment. Additionally, this work illustrates that the attention module can yield significant benefits without substantially increasing computational costs. Extensive experiments demonstrate that the proposed network holds a fundamental advantage over baseline Transformers, achieving state-of-the-art performance on two benchmark datasets. The proposed method anticipate that our approach will stimulate further research in 2D to 3D pose lifting, considering various ambiguities.

However, the proposed model faces challenges that need to be considered in future work. Firstly, training and predicting occluded joints proved to be difficult for the architecture. Implementing techniques to handle the hypothesis of 3D Pose could address this issue. Secondly, the computational demands of end-to-end networks pose a hurdle for real-time applications due to their significant computational load. In future research, this paper aims to mitigate this computational cost and develop a lightweight system.

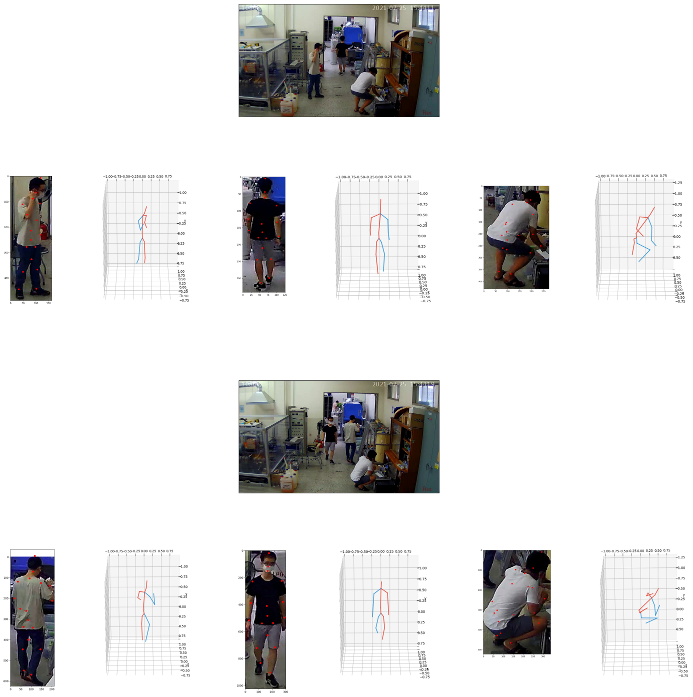

Fig. 5. Qualitative result for 3D human pose estimation in frame number 1308 and frame number 1469 of ISLAB industrial dataset - video 1

References

[1] Sania Zahan, Ghulam Mubashar Hassan, Ajmal Mian, *SDFA: Structure-Aware Discriminative Feature Aggregation for Efficient Human Fall Detection in Video*, *IEEE Transactions on Industrial Informatics*, vol. 19, no. 8, pp. 8713-8721, 2023, doi: `https://doi.org/10.1109/TII.2022.3221208`.

[2] Michał Wieczorek, Jakub Siłka, Marcin Woźniak, Sahil Garg, Mohammad Mehedi Hassan, *Lightweight Convolutional Neural Network Model for Human Face Detection in Risk Situations*, *IEEE Transactions on Industrial Informatics*, vol. 18, no. 7, pp. 4820-4829, 2022, doi: `https://doi.org/10.1109/TII.2021.3129629`.

[3] Hai Liu, Tingting Liu, Zhaoli Zhang, Arun Kumar Sangaiah, Bing Yang, Youfu Li, *ARHPE: Asymmetric Relation-Aware Representation Learning for Head Pose Estimation in Industrial Human–Computer Interaction*, *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7107-7117, 2022, doi: `https://doi.org/10.1109/TII.2022.3143605`.

[4] Nurul Amin Choudhury and Badal Soni, *An Adaptive Batch Size-Based-CNN-LSTM Framework for Human Activity Recognition in Uncontrolled Environment*, *IEEE Transactions on Industrial Informatics*, vol. 19, no. 10, pp. 10379-10387, 2023, doi: `https://doi.org/10.1109/TII.2022.3229522`.

[5] Mohammed A. A. Al-qaness, Abdelghani Dahou, Mohamed Abd Elaziz, A. M. Helmi, *Multi-ResAtt: Multilevel Residual Network With Attention for Human Activity Recognition Using Wearable Sensors*, *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 144-152, 2023, doi: `https://doi.org/10.1109/TII.2022.3165875`.

[6] Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, Qiang Fu, *Robotic Continuous Grasping System by Shape Transformer-Guided Multiobject Category-Level 6-D Pose Estimation*, *IEEE Transactions on Industrial Informatics*, vol. 19, no. 11, pp. 11171-11181, 2023, doi: 10.1109/TII.2023.3244348.

[7] Xinjian Deng, Jianhua Liu, Honghui Gong, Hao Gong, Jiayu Huang, *A Human–Robot Collaboration Method Using a Pose Estimation Network for Robot Learning of Assembly Manipulation Trajectories From Demonstration Videos*, *IEEE Transactions on Industrial Informatics*, vol. 19, no. 5, pp. 7160-7168, 2023, doi: `https://doi.org/10.1109/TII.2022.3224966`.

[8] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, Jingdong Wang. *HRFormer: High-Resolution Transformer for Dense Prediction*. In *NeurIPS*, 2021.

[9] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2875–2882.

[10] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2014.

[11] Alejandro Newell, Kaiyu Yang, Jia Deng, *Stacked Hourglass Networks for Human Pose Estimation*, in *European Conference on Computer Vision (ECCV)*, 2016, pages=483–499, organization=Springer.

[12] Xue, Youze and Chen, Jiansheng and Gu, Xiangming and Ma, Huimin and Ma, Hongbing, “Boosting Monocular 3D Human Pose Estimation With Part Aware Attention,” in IEEE Transactions on Image Processing, vol.31, pp.4278-4291, June .2022, doi=10.1109/TIP.2022.3182269

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need.” Advances in Neural Information Processing Systems, pp. 5998-6008, 2017.

[14] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool, MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation, 2022, `https://arxiv.org/abs/2111.12707`, arXiv:2111.12707 [cs.CV].

[15] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu, AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 7157-7173, 2023, doi: 10.1109/TPAMI.2022.3222784.

[16] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little, A simple yet effective baseline for 3D human pose estimation, 2017, `https://arxiv.org/abs/1705.03098`, arXiv:1705.03098 [cs.CV].

[17] Tong Zhang, Jingxiang Lian, Jingtao Wen, C. L. Philip Chen, *Multi-Person Pose Estimation in the Wild: Using Adversarial Method to Train a Top-Down Pose Estimation Network*, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 7, pp. 3919-3929, 2023, doi: `https://doi.org/10.1109/TSMC.2023.3234611`.

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. 2019. arXiv:1902.09212 [cs.CV].

[19] Tianhan Xu and Wataru Takano. Graph Stacked Hourglass Networks for 3D Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 16105–16114, 2021. DOI: `10.1109/CVPR.2021.00161`

[20] MPI-INF-3DHP Dataset, Max Planck Institute for Informatics, `http://gvv.mpi-inf.mpg.de/3dhp-dataset/`,

[21] Shuangjun Yang, Huan Li, Yihui Li, Jiaying Wang, Hao Wang, and Hongsheng Li. PoseFormer: Generalized 3D Human Pose Estimation in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 945-954, 2021. DOI: `10.1109/CVPR42942.2021.00096`

[22] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning Pose Grammar to Encode Human Body Configuration for 3D Pose Estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[23] 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6173-6183, 2020. DOI: `10.1109/CVPR42600.2020.00617`

[24] Arindam Das, Sudip Das, Ganesh Sistu, Jonathan Horgan, Ujjwal Bhattacharya, Edward Jones, Martin Glavin, Ciarán Eising, Deep Multi-Task Networks For Occluded Pedestrian Pose Estimation, 2022, arXiv preprint arXiv:2206.07510, primaryClass=cs.CV

[25] Siyuan Li, Zedong Wang, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, Stan Z. Li, Efficient Multi-order Gated Aggregation Network, 2023, arXiv preprint arXiv:2211.03295, primaryClass=cs.CV

[26] Tien-Dat Tran, Xuan-Thuy Vo, Ashraf Russo, and Kang-Hyun Jo, Simple Fine-Tuning Attention Modules for Human Pose Estimation, in Proceedings of the Conference Name, November 2020, pp. 175-185, ISBN: 978-3-030-63118-5, doi: 10.1007/978-3-030-63119-215.

[27] Xiao Wei, Hao-Shu Hsu, Chu-Song Huang, Xiaoou Tang, Simple Baselines for Human Pose Estimation and Tracking, in *European Conference on Computer Vision (ECCV)*, 2018, pp. 466–481, doi: 10.1007/978-3-030-01246-5_29.

[28] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, Jiebo Luo, *Anatomy-aware 3D Human Pose Estimation with Bone-based Pose Decomposition*, *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[29] Dario Pavllo, Christoph Feichtenhofer, David Grangier, Michael Auli, *3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages=7753–7762, 2019.

[30] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, Ross Girshick, *Detectron2*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pages=3964–3973.