Improved 2D-3D Human Pose Estimation System with Proposed Attention Mechanism

***Abstract*— Neural networks have achieved significant success in lifting 2D to 3D pose estimation. However, effectively minimizing the redundant 2D pose sequences from a weak pose detector continues to be a challenging issue. To tackle this, the proposed method incorporates the attention mechanisms inside the proposed system. This system comprises two main components: a 2D pose detector and a 3D pose estimator. The 2D pose detector is enhanced with a new attention module called channel attention, which is implemented after the last two blocks to improve accuracy. Additionally, the 3D pose network utilizes a Transformer-based architecture with advanced attention mechanisms. It introduces a new Transformer Encoder that applies spatial and temporal attention to capture important information in 2D pose sequences. This proposed architecture shows promising comparative performance on two benchmark datasets for 3D human pose estimation Human3.6M and MPI-INF-3DHP. Moreover, the 3D Network improves performance by 0.9% and 0.3%, respectively over its closest counterpart, PoseFormer. Furthermore, in terms of 2D pose estimation, the system outperforms existing methods on the COCO 2017 Microsoft Dataset. Link demo: demo vision**

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

I. Introduction

A. Research Background

THREE dimensional (3D) Human Pose Estimation (HPE) is a critical area of study in computer vision. This technique aims to determine the three-dimensional coordinates of human body joints from a two-dimensional image or a series of images. Human pose estimation has a variety of applications, including object recognition [1], [2], human-computer interaction [3], activity recognition [4], [5], and robotic systems [6], [7].

1) 2D Human Pose Network: In the field of 2D Human Pose Estimation, as outlined in the introduction, most techniques fall into two main categories: top-down and bottom-up. Recently, bottom-up methods [8] have become popular due to their efficiency. These methods predict keypoints directly from the input image without requiring person detection. However, because they do not focus specifically on human regions, their accuracy may be compromised. Conversely, top-down methods start with a human detector that identifies all individuals in an image and then performs single-person pose estimation for each detected subject, resulting in more accurate predictions. Notable techniques in this category include HRNet [9] and HRFormer [10]. This paper introduces a novel top-down approach that significantly enhances heatmap prediction by applying an attention mechanism between the characteristic functions of the predicted and ground truth (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 methods. One-stage approaches directly infer 3D poses from input images without intermediate 2D pose representations [11], [12], while two-stage network 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 HPE, this 2D-to-3D pose lifting method can efficiently and accurately regress 3D poses using detected 2D key points. Despite the promising results achieved by using temporal correlations from fully convolutional [4], [13] 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 [15], the 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. [16] 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 the 2D Pose Estimator, deep convolutional neural networks have demonstrated exceptional performance. Typically, most existing approaches process the input through a network to enhance the resolution and subsequently apply 3D Human Pose Estimation (HPE) on the 2D results, as depicted in Fig. 1. The 3D network, which uses a series of 2D points as input, generally consists of high-to-low resolution subnetworks arranged in sequence. For instance, the Hoursglass model [17] deploys a symmetric low-to-high resolution technique to regain high resolution, while Simple Baseline [8] utilizes a few transposed convolution layers to create high-resolution representations. Nevertheless, accurately lifting the 2D keypoints to a 3D model remains a significant challenge.

Recent advancements in 3D human posture encoding have been facilitated by deep neural networks [18], [19]. However, these networks encounter several challenges. First, improving the accuracy of various network types, such as real-time networks or networks that measure accuracy, is crucial. Second,

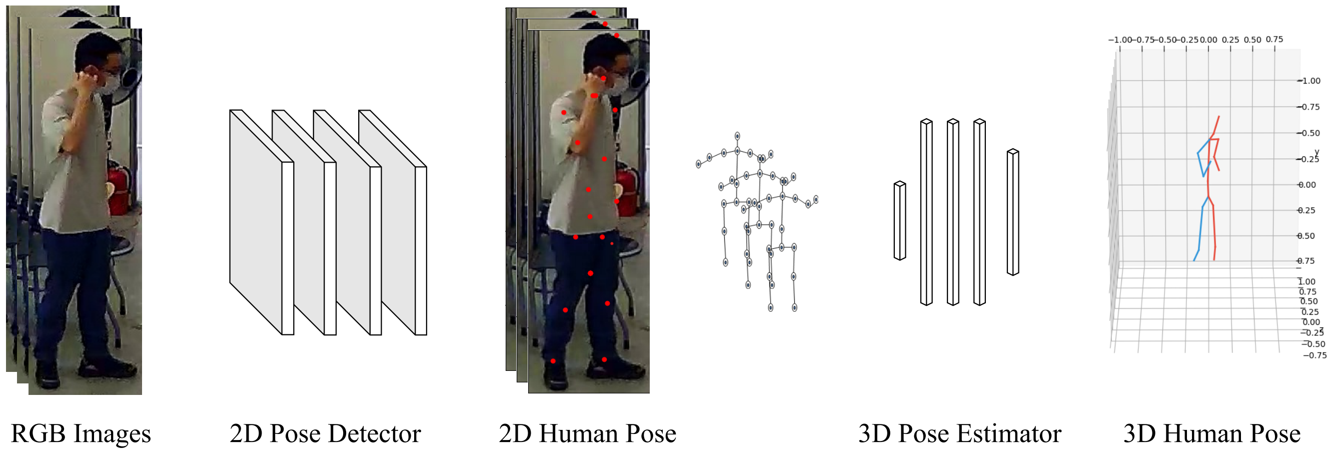

Fig. 1. The proposed system comprises two main components: the 2D Pose Detector and the 3D Pose Estimator. The 2D Pose Detector processes the input image to identify 2D human keypoints. Subsequently, the 3D Pose Estimator takes a sequence of these predicted 2D joints from the Detector and accurately estimates the final 3D pose of the human figure.

it is common practice to verify the accuracy of a network by using different 2D pose results. Finally, the current challenge for networks is to enhance accuracy while either maintaining or increasing processing speed. The proposed study introduces a novel network structure and evaluates it in terms of speed and accuracy. This experiment diverges from PoseFormer [15] by implementing a new attention mechanism known as spatial-temporal attention.

C. Attention in Human Pose Estimation Review

The attention mechanism has been widely adopted in natural language processing (NLP) tasks, achieving state-of-the-art performance in machine translation [5] and language understanding [15]. Recently, attention-aware features have also proved highly effective in computer vision tasks. For instance, Newel et al. [17] proposed a robust attention module that integrates an attention branch with an hourglass block, which is stacked multiple times to construct a deep convolutional neural network for image classification. Leveraging the self-attention mechanism, the network described in [20] captures rich contextual dependencies for scene segmentation. Similarly, Zhang et al. [18] and Yang et al. [15] have incorporated attention mechanisms into various convolutional neural networks to enhance human pose estimation. A prominent mechanism in this area is self-attention, also known as transformer-based attention, which enables the model to focus on different parts of the input and recognize long-range dependencies. This capability allows pose estimation models to dynamically prioritize the significance of different joints or body parts based on their interrelations.

Furthermore, spatial attention can be utilized to emphasize relevant spatial regions within an image, enhancing the model's focus on crucial areas for accurate pose estimation through technologies like spatial transformer networks or spatial attention modules.

D. Contribution of The Paper

Over the past few years, there has been a significant increase in research focused on 2D and 3D human pose estimation. 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 introduces and applies a new attention mechanism to both the 2D pose detector and the 3D estimator, enhancing the network's ability to improve occlusion issues. Inside the 2D Pose Network, a new channel attention module deploying 1×1 depth-wise convolutions across different channels effectively captures the important keypoint information. Additionally, a new spatial-temporal attention mechanism was implemented in the 3D Network, significantly increasing the accuracy of 3D predictions.
2) The study presents a comprehensive system for Lifting 2D-3D Pose Estimation. The proposed architecture accurately predicts the final 3D human posture from the input image, incorporating several minor techniques to enhance both 2D and 3D results.
3) The proposed method, which is straightforward and has not increased much in computational cost, surpasses the compared methods in performance on benchmark datasets. For 2D, it is extensively compared with other methods on the Microsoft COCO 2017 benchmark. Additionally, this method achieves competitive results on the Human3.6M and MPI-INF-3DHP datasets for the 3D Network.

II. Methodology

A. 2D Pose Estimator

[ht]

1) Backbone network: The proposed system utilizes a benchmark composed of HighResolutionNet-W32 and HighResolutionNet-W48 [9], as depicted in Fig. 2. Each HighResolutionNet can be organized in two ways: three subnetworks or three stages that include residual blocks, a

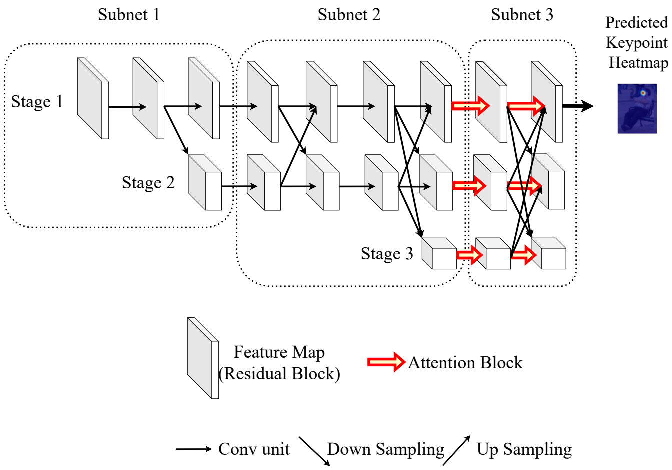

Fig. 2. The proposed 2D Pose Detector architecture incorporates a multi-resolution framework adapted from the original HRNet, which includes both downsampling and upsampling processes. The key modification in the proposed design is the integration of the attention mechanism inside the third sub-network, enhancing the model's focus and performance in key areas.

convolution unit, and 3×3 convolution for downsampling, along with bi-linear interpolation for upsampling. The default input image is resized to dimensions of 256×192 for both HighResolutionNet-W32 and HighResolutionNet-W48 models. For each stage inside the network, the feature map with the initial dimensions of $H \times W$ is halved, and the channel count C doubles after every stage. Consequently, by the end of the backbone, the feature map size is reduced to $\frac{W}{4} \times \frac{H}{4} \times 4C$. The architecture is called HR-Net because it maintains the dimension of $W \times H$ from the beginning until the regression process. Additionally, the numbers 32 or 48 in the backbone network's name refer to the number of channels, which increase to 128 and 192 at the final stage, respectively. The network utilizes mean square error loss, as introduced in Section II.A.3, to generate the predicted keypoint from ground truth.

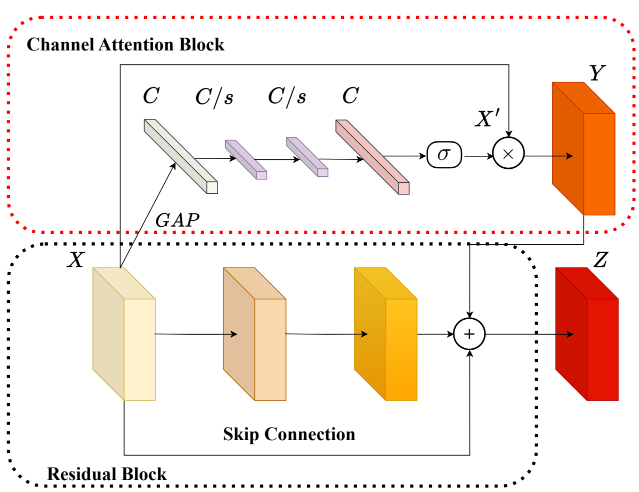

Fig. 3. Architecture of Attention module. The Attention was implemented on the last layer of the residual block

2) Attention Module: In the proposed 2D Pose Estimator, the Attention Mechanism was applied only to the last two blocks of the final sub-network. As shown in Fig. 2, only six attention modules were deployed to balance computational cost and accuracy. According to Fig. 3, the attention module deployed is based on channel attention, no spatial attention was used because of inefficient for keypoints. The reason is Spatial tries to apply Global-pooling or Max-pooling to get the global context but it also makes the keypoint information collapse. Hence, the proposed architecture believes that only channel attention is better than used both channel and spatial. After one residual block in the backbone network, the feature information is directed to the channel attention module where convolutions $Conv_{1\times1}$ are applied. The output of the Channel Attention module is:

$$X' = \sigma(Conv_{1\times1}(Conv_{1\times1}(Conv_{1\times1}(GAP(X))) \quad (1)$$

where $X \in \mathbb{R}^{W\times H\times C}$ is the input of the feature map from the previous residual block of the Channel Attention Block and $X' \in \mathbb{R}^{1\times1\times C}$ is the output of attention map. $Conv_{1\times1}$ include 1×1 convolution layer, batch normalization, and RELU (Rectified Linear Unit) for activation function. GAP is the global average pooling to capture the channel information. σ is the sigmoid function to generate the probability of the attention area. The tensor information in the Channel Attention Module (CAM) uses this convolution to reduce the channel dimension from C to $\frac{C}{s}$. s is the shrinking ratio, which is typically set to 4. After that, a multi-element-wise $\bigotimes$ function is applied to generate attention feature Y:

$$Y = X \bigotimes X', \quad (2)$$

Finally, combining the Channel Attention Block output with the Residual Block by addition-element-wise $\bigoplus$

$$Z = Conv_{3\times3}(Conv_{3\times3}(X)) \bigoplus X \bigoplus Y \quad (3)$$

where architecture of $Conv_{3\times3}$ is the same with $Conv_{1\times1}$ but utilizes the convolution layers 3×3. First $\bigoplus$ refers to the skip connection inside the residual block, and second $\bigoplus$ combines the residual block and channel attention module. Additionally, the size of the feature map kept the same size for $X, Y, Z \in \mathbb{R}^{H\times W\times C}$

3) 2D Pose Estimator Loss: Heat maps are utilized in the proposed work to demonstrate body keypoint locations in the loss function. The principles of Ground-truth heat map H_n is then built up by utilizing the Gaussian distribution and the mean a_n with variance τ as illustrated in the next equation.

$$H_n(p) \sim N(a_n, \tau), \quad (4)$$

where $\mathbf{p} \in \mathbb{R}^2$ illustrate the coordinate, and τ is an identity matrix $\mathbf{I}$. The final layer of the proposed architecture generated A heat maps, *i.e.*, $\hat{S} = \hat{S}_b^a$ and $b = 1^B$ for B body joints. The mean square error for the loss function is defined, which is summarized as follows:

$$L = \frac{1}{AB} \sum_{A=1}^{A} \sum_{B=1}^{B} \left\| S_b^a - \hat{S}_b^a \right\|^2, \quad (5)$$

where A also denotes the number of selected in the training process, B denotes the number of joints. S_b^a and $\hat{S}_b^a$ is the predict and ground truth for 2D Keypoint.

B. 3D Pose Estimation Network

1) Baseline network: In this work, it adopt a Transformer-based architecture which is in Fig. 4 since it performs well in long-range dependency modeling. Then first give a brief description of the basic components in the Transformer [20], including a multi-head self-attention (MSA) and a multi-layer perceptron (MLP). 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:

$$MSA(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_m}})V, \tag{6}$$

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:

$$MLP(x) = \alpha(xW_1 + a_1)W_2 + a_2, \tag{7}$$

where α denotes the Gaussian Error Linear Unit (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. Layer Norm (LN) is used before for all MSA and MLP to make the network balance between accurate and computational cost.

2) Spatial Transformer: The Spatial Transformer tries to catch the information inside the individual pose. Hence this paper proposed a new Spatial Attention (SA) module to focus on each key point by group of 5. This module is inserted between the LN layer and MLP for $N_1\times$ transformer block. The Spatial attention module consists of two depth-wise convolutions with kernel size 5, group normalization, and non-linearity GELU. Also, the skip 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:

$$P_0 = Conv(Norm(GELU(Conv(P)))) + P, \tag{8}$$

where $GELU$ refers to the non-linear layer, $Conv$ is the standard convolution layer with kernel 1×5 and Norm indicates the normalization used in [15]. Since the focus of the SA module is on the interaction between the joints, the output of the MSA part in Eq. 6 has been transposed. That is, it becomes $P_0 \in R^D$. The spatial encoders for a transformer layer can be represented by the following list of operations:

$$MLP(x_0) = \beta(xW_1 + a_1)W_2 + a_2, \tag{9}$$

where β denotes the P function in Eq. 8

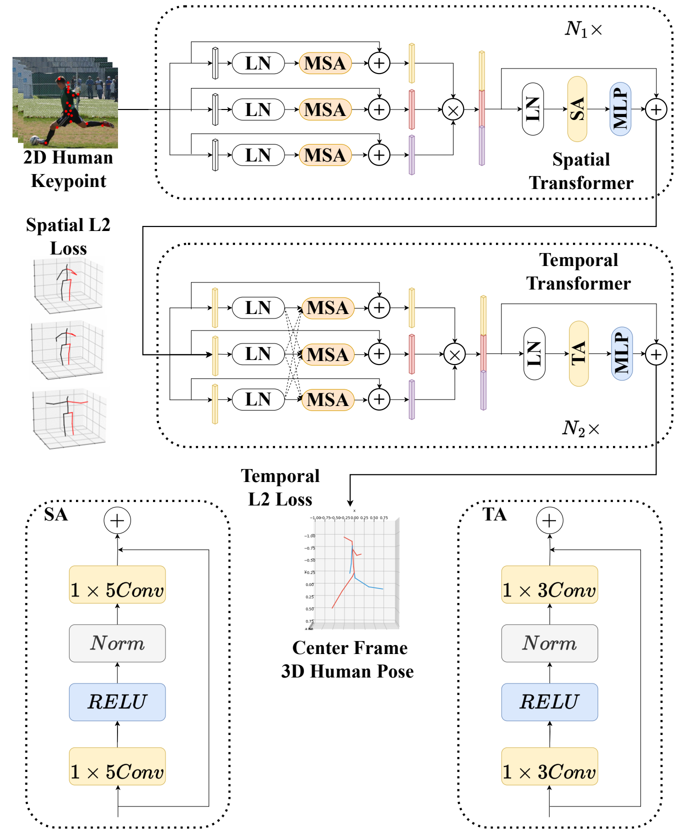

Fig. 4. Detailed Architecture of 3D Pose Estimator. The proposed network is based on the transformer. The new thing here is the 1×1 conv with different channels for Spatial and Temporal attention. In the case of Temporal attention, the attention-getting interaction among multi-feature

3) Temporal Transformer: Same with the Spatial Transformer, The Temporal attention (TA) is applied for the $N_2\times$ transformer blocks. Inside the temporal transformer, it 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 $P_1 \in R^{F \times J \times D}$ and then using convolutions this paper extract K, Q, and V such that:

$$K = PW_k, Q = PW_q, V = PW_v \tag{10}$$

The TA module is same as SA, but the difference is the Conv kernel size is 1×3. The output embedding P_1 and $MLP(x_1)$ can defined:

$$P_1 = Conv(Norm(GELU(Conv(P)))) + P, \tag{11}$$

$$MLP(x_1) = \beta(xW_1 + a_1)W_2 + a_2, \tag{12}$$

4) Regression Head: To understand clearly, suppose the 3D network input is the 2D Pose Sequence $X_{2D} \in \mathbb{R}^{N \times J \times 2}$ which N is the number of frame, J is the number of keypoint and 2 is mean the 2D coordinate of Keypoint (x, y). Then X_{2D} go to Spatial Transformer and it outputs the $(X')_{3D} \in \mathbb{R}^{N \times J \times 3}$ for each X_{2D} respectively. After that, X'_{3D} is also the input of Temporal Transformer to generate $Z_{L3} \in \mathbb{R}^{N \times J \times 3}$ and put it to the regression.

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 Ours final prediction, 3 here is the 3D coordinate (x, y, z)

5) 3D Loss: The entire 3D Estimator is trained in an end-to-end manner with a Mean Square Error loss for the Spatial module function defined by the mean of MPJPE, which is calculated as follows:

$$\mathcal{L}_s = \sum_{m=1}^{M} \sum_{j=1}^{J} \left\| S_j^m - \hat{S}_j^m \right\|_2, \tag{13}$$

Where *M* denotes the number of selected 2D Pose in the training process, *J* is the number of Joint. S_j^m is the predict 3D human pose joint and $\hat{S}_j^m$ is the ground truth 3D Pose. Same with the Spatial L2 Loss, The Temporal L2 Loss is calculated as follows:

$$\mathcal{L}_t = \sum_{N=1}^{N} \sum_{j=1}^{J} \left\| S_j^n - \hat{S}_j^n \right\|_2, \tag{14}$$

Where *N* denotes the number of selected predicted 3D Pose in the training process. The total loss for 3D network composed in:

$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_t \mathcal{L}_t, \tag{15}$$

While λ_s and λ_t is the weighted parameter for each loss

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.

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 has 17 key-point 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 evaluates proposed model on two general datasets: Human3.6M [24] and MPI-INF-3DHP [25].

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 [15], the proposed network adopts the MPJPE (P1), percentage of correct keypoints (PCK) with 150 mm, and area under the curve (AUC) results as the evaluation metrics.

B. Implementation Details

The proposed model, implemented using PyTorch, utilizes 2D keypoints from HRNet [9], CPN Detector, or 2D ground truth to analyze performance. The 2D pose detector in this study is based on the AlphaPose [22] codebase, while the 3D pose estimator adopts the PoseFormer codebase [15]. Although the proposed model is capable of adapting to any length of the input sequence, for fairness in comparison, specific sequence lengths (T) were chosen for three datasets: Human3.6M (T = 81, 243), and MPI-INF-3DHP (T = 1, 27). Details regarding the selection of frame lengths are discussed in the ablation study (Section III.E.3). The batch size, dropout rate, and activation function are set at 1024, 0.1, and GELU, respectively. All experiments were conducted on the PyTorch framework using two NVIDIA GeForce GTX 2080 Ti GPUs. The network training deploys the Adam optimizer [26], with a learning rate of 0.001 and a decay factor of 0.95 applied every two epochs.

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. To ensure a fair comparison, we evaluated the results against networks without pretraining. Despite being only trained on COCO, the proposed network still surpasses the ImageNet-pretrained HRNet-W32 and HRNet-W48 by 1.3% and 1.6% in Average Precision (AP), respectively. This demonstrates that the integration of the attention mechanism can outperform models that rely on pretraining.

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. To guarantee fair comparisons, the input is taken from CPN in the form of 2D keypoints for training and testing. Table II 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

TABLE I
COMPARISON RESULT ON COCO 2017 VALIDATION SET. PT = PRETRAIN THE BACKBONE ON THE IMAGENET CLASSIFICATION TASK

Methodology	Backbone	PT	#Parameters (M)	Image size	AP (%)	AR (%)	AP^{50}	AP^{75}	AP^{L}	AP^{M}
Fine-tune Attention [27]	ResNet-50	N	31.2M	256×192	71.4	76.3	91.6	78.6	75.7	68.2
Fine-tune Attention [27]	ResNet-101	N	50.2M	256×192	72.3	77.1	92.0	79.4	77.1	68.3
High-Resolution Net [9]	HRNet-W32	N	28.5M	256×192	73.4	78.9	89.5	80.7	80.1	70.2
High-Resolution Net [9]	HRNet-W32	Y	28.5M	256×192	74.4	79.8	90.5	81.9	81.0	70.8
High-Resolution Net [9]	HRNet-W48	Y	63.6M	256×192	75.1	80.4	90.6	82.2	81.8	71.5
Zhang at al. [18]	HRNet-W32	N	29.2M	256×192	74.8	77.6	92.5	81.6	79.3	72.0
Zhang at al. [18]	Hourglass-8	N	25.8M	256×192	75.1	80.4	90.6	82.6	81.9	71.6
MogaNet-T [16]	MogaNet	N	8.1M	256×192	73.2	81.0	90.1	78.8	-	-
MogaNet-S [16]	MogaNet	N	29M	256×192	74.9	82.8	90.7	80.1	-	-
PPE-Net [28]	ResNeXt-101	Y	-	256×192	75.7	-	90.3	76.3	79.5	80.7
Ours	HRNet-W32	N	29.3M	256×192	75.5	80.5	90.4	82.0	82.2	71.3
Ours	HRNet-W48	N	65.9M	256×192	76.0	80.7	90.6	82.6	82.9	71.8

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.

Protocol # 1 - CPN	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Punch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [23]	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.* [19] ∗	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.* [13]	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 [24]	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.* [11]	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.* [15] †	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**
Ours	**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.* [19] ∗	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
Pavllo *et al.* [12] ⋆	34.1	36.1	34.4	**37.2**	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Wu *et al.* [29]	**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.* [15] †	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.* [13]	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
Ours	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

⋆ denotes that the 2D Keypoint detection is the cascaded pyramid network(CPN).
† refers to 3D network apply transformer-based model.
∗ refers to used 2D Groundtruth for training
Bold: is the best performance

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 - GroundTruth	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Punch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [23]	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.* [19]	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.* [13]	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 [24]	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.* [11]	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.* [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32.3
Yang *et al.* [15] †	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
Ours	**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**

† refers to 3D network apply transformer-based model.

worth noting that the across-joint modules in the spatial and temporal cases are crucial to infer the body-joint 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 achieves 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 the proposed 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 a 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 the baseline PoseFormer 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 among the compared approaches.

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 (mm) ↓
Pavllo *et al.* [12] (f = 81)	86.0	51.9	84.0
Lin *et al.* [20] (f = 25)	83.6	51.4	79.8
Li *et al.* [13]	81.2	46.1	99.7
Chen *et al.* [30]	87.9	54.0	78.8
Yang *et al* [15] (f = 9)	88.6	56.4	75.5
Ours (f = 9)	89.1	57.5	76.3

↑ Highest result is the best.
↓ Lowest result is the best.

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. 5 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 (mm) ↓
CPN	47.6	37.4	3.20
Detectron2 [29]	45.7	37.0	3.02
Hoursglass [17]	52.3	41.2	4.11
HRNet-W32 [9]	45.1	36.3	2.91
HRNet-W32+AM (Ours)	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 compared to the baseline.

TABLE VI
COMPARISION RESULT OF EACH MODULE IN THE PROPOSED SYSTEM

Method	2D Attention	3D SAM	3D TAM	MPJPE (mm) ↓
PoseFormer				44.3
Ours	✓			43.6
Ours		✓		43.7
Ours			✓	43.8
Ours		✓	✓	43.3
Ours	✓	✓	✓	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 (16 Attention modules got added) got the best result however it also got the highest number of parameters in the computational cost. Besides, Table VII 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 sub-network and stage (only 6 were added) to not only balance the computational cost but also keep the high accuracy.

TABLE VII
THE RESULT WHEN UTILIZING THE ATTENTION MECHANISM FOR EACH SUB-NETWORK AND EACH STAGE OF HRNET-W32

Backbone	Sub-Net	AP (%)	#Param (M)
HRnet-W32	-	73.4	28.5M
HRnet-W32	1	74.2	28.7M
HRnet-W32	2+1	74.6	28.9M
HRnet-W32	3+2+1	75.5	29.3M
Backbone	Stage	AP	#Param
HRnet-W32	1	74.3	28.9M
HRnet-W32	2+1	74.8	29.1M
HRnet-W32	3+2+1	75.5	29.3M

Table VIII showcases the influence of different positions of the SAM and TAM on 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 decided to put SAM and TAM between the MSA and MLP.

TABLE VIII
THE RESULT WHEN APPLYING DIFFERENT POSITIONS OF 1×1 CONVOLUTION IN SAM AND TAM

Module	Before MSA	After MSA	After MLP	MPJPE (mm) ↓
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 [12], PoseFormer by PoseFormer et al. [15], 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 ("Ours") 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) ↓
Pavllo *et al.* [12]	243 (default)	12.75M	47.5
Pavllo *et al.* [12]	81	12.70M	47.9
Yang *et al.* [15]	81 (default)	9.59M	44.3
Yang *et al.* [15]	27	9.60M	44.6
Ours	9	9.85M	44.3
Ours	27	9.86M	43.6
Ours	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	Learning rate	#Param (M)	MPJPE (mm) ↓
Pavllo *et al.* [12]	0.25 (default)	12.70M	47.9
Pavllo *et al.* [12]	0.1	12.70M	47.5
Yang *et al.* [15]	0.25 (default)	9.60M	44.3
Yang *et al.* [15]	0.1	9.60M	44.6
Ours	0.25	9.86M	43.3
Ours	0.2	9.86M	43.3
Ours	0.1	9.86M	43.1
Ours	0.05	9.86M	43.4

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 Ours 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 Ours approach will stimulate further research in 2D to 3D pose lifting, considering various ambiguities. In future research, this paper aims to mitigate this computational cost and develop a lightweight system.

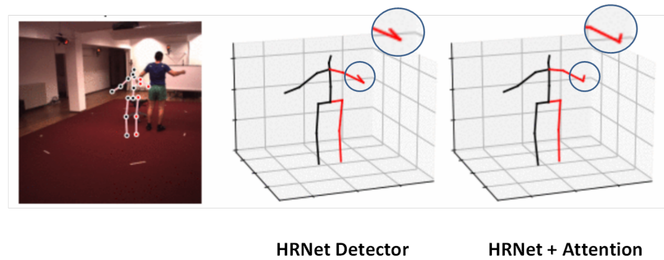

Fig. 5. 3D human pose estimation result come from 2D skeleton based on detector and detector with attention mechanism

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] 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.

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

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

[11] 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

[12] 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.

[13] 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

[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] 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

[16] 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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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

[22] 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.

[23] 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].

[24] 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.

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

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

[27] 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.

[28] 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

[29] 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.

[30] 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.