Group Spatial Attention for 3D Human Pose Estimation

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

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

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

I. Introduction

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

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

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

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

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

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

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

II. Related work

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

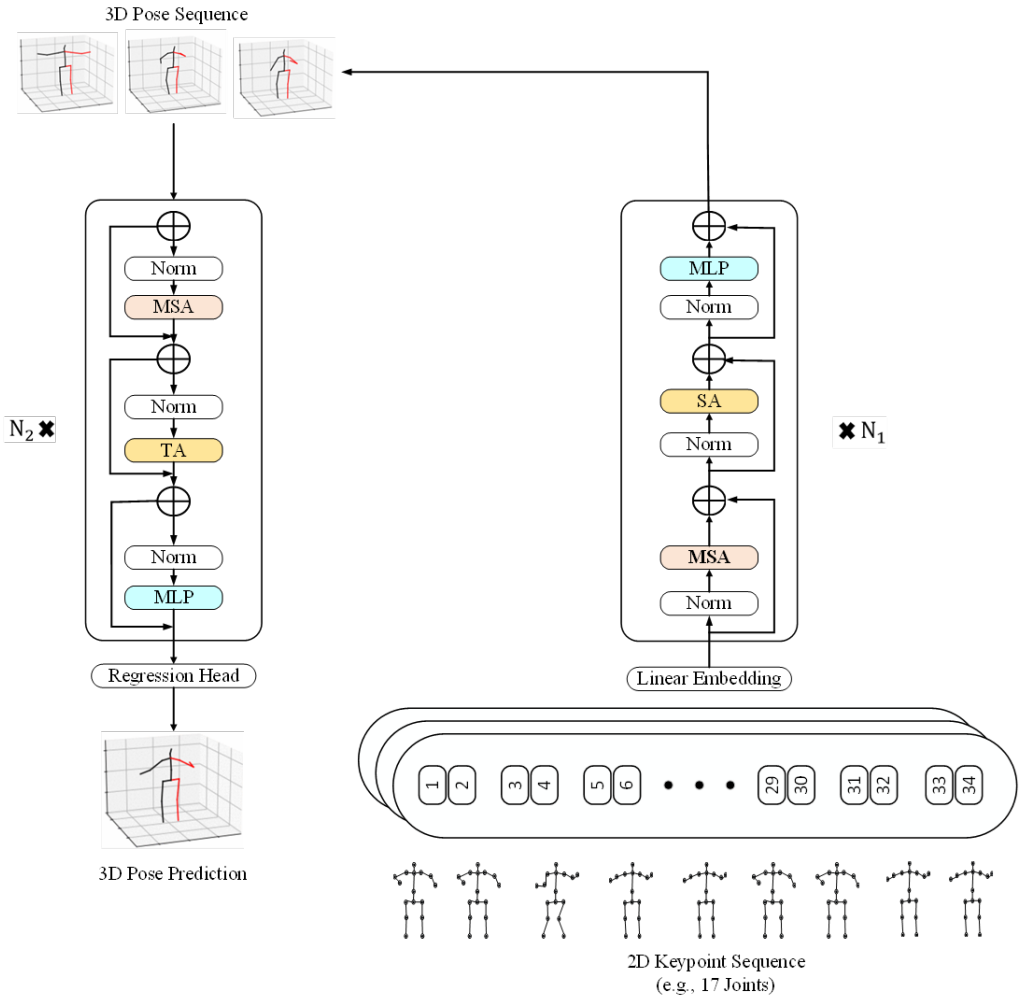

Fig. 1. Illustrating the architecture of the proposed 3D-human-pose estimator. The proposed network training with the 2D GrouthTruth and 2D information from HRNet

In the proposed paper, we apply the top-down method for the whole architecture which is illustrated in Fig.2, From the input images, the model utilizes the existing detector for human detection. YOLO [**?**] is of diversity kind of detector, which has many versions for different cases such as real-time, high accuracy, or for mobile devices. To balance everything, the proposed method utilizes the YOLO-V3. After applying the detector to the human region, the whole network utilizes the pose estimator to perform training tasks in the human region. Additionally, data augmentation will apply in this stage. In comparison, the top-down strategy employs enough viewpoint for implementing a network, which makes the network increase the accuracy but lose the sufficient speed

3D Pose Estimation: 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 [3], [4], 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 [**?**], [1] 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 [5]. Following PoseFormer [6], 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. [7] 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.

III. METHODOLOGY

A. 3D Pose Estimation Network

1) Baseline network: This work adopts a Transformer-based architecture, depicted in Fig. 4, known for its robust performance in modeling long-range dependencies. Initially, we briefly describe the core components of the Transformer, as introduced in [8], which include a multi-head self-attention (MSA) and a multi-layer perceptron (MLP). The inputs $x \in \mathbb{R}^{n\times d}$ are linearly projected 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 represents the sequence length and d the dimensionality. The scaled dot-product attention is computed as:

$$MSA(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_m}}\right)V, \tag{1}$$

where Q, K, and V are split into h heads and processed in parallel, and the results are concatenated. The MLP includes two linear layers for non-linear transformation and feature processing:

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

with $W_1 \in \mathbb{R}^{d\times d_m}$ and $W_2 \in \mathbb{R}^{d_m\times d}$ being the weights, and $a_1 \in \mathbb{R}^{d_m}$ and $a_2 \in \mathbb{R}^d$ the biases. A Layer Normalization (LN) precedes both MSA and MLP to optimize accuracy and computational efficiency.

2) Spatial Transformer: The Spatial Transformer captures detailed pose information through a new Spatial Attention (SA) module, focusing on groups of five keypoints. This module, embedded between the LN and the MLP within the $N_1\times$ transformer block, employs two depth-wise convolutions with a kernel size of 5, group normalization, and GELU activation. A skip connection is also included to prevent overfitting. The transformations applied to the output of the patch embedding step P_0 are given by:

$$P_0 = Conv(Norm(GELU(Conv(P)))) + P, \tag{3}$$

where $Conv$ applies a 1×5 kernel and $Norm$ represents the normalization process referenced in [6]. The spatial encoders in a transformer layer are represented by:

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

3) Temporal Transformer: Similarly, the Temporal Transformer (TA) within the $N_2\times$ transformer blocks captures temporal dynamics by learning pairwise feature correlations through an outer product mechanism. Each element of the correlation matrix $C_{ij} = \sum_F P_i P_j$ represents the dot product of embedded features from frames i and j, pooled by summation, where $P_i \in \mathbb{R}^{J\times D}$ are the features of frame i. This process involves a transformation that combines positional information with the frame features:

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

The TA module operates similarly to the SA module but with a convolution kernel size of 1×3. The processed embeddings and MLP transformations are defined by:

$$P_1 = Conv(Norm(GELU(Conv(P)))) + P, \tag{6}$$

B. Loss Function

The Proposed network utilizes Heat maps to demonstrate body Keypoint locations in whole Loss Function. In Fig. 3 the GrouthTruth coordinate by $a = \{a_k\}\, k = 1^K$, where $a_k = (x_k, y_k)$ is the spatial position of the kth keypoiny in the trained sample. The heat map value H_k of groundtruth is then constructed after applying the Gaussian function with variance $\sum$ and the mean a_k as shown below.

$$H_k(p) \sim N(a_k, \sum) \tag{7}$$

where $\mathbf{p} \in \mathbb{R}^{\mathbf{2}}$ illustrate the coordinate, and $\sum$ is experimentally defined as an identity matrix $\mathbf{I}$. In the final process of training, the network will predict K heat maps, *i.e.,* $\hat{S} = \left\{\hat{S}_k\right\} k = 1^K$ for K body joints. Mean Square error is the main Loss, which is calculated as follows:

$$L = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}\left\|S_k - \hat{S}_k\right\|^2 \tag{8}$$

N denotes the total of images in the training process. Using information from the backbone network's last layer, The proposed architecture generated the predicted heatmap keypoint by using the ground truth.

IV. EXPERIMENTS

A. Datasets and Evaluation Protocols

About the 3D human pose, this approach evaluate proposed model on two general datasets: Human3.6M [9], MPI-INF-3DHP [10] and Industrial dataset individually.

1) Human3.6M: is the most commonly used indoor dataset for the 3D human pose estimation tasks. Following the same policy of the base method [5], 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.

2) 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 [6], the proposed network adopts the MPJPE (P1), percentage of correct keypoints (PCK) with 150mm, and area under the curve (AUC) results as the evaluation metrics.

B. Implementation Details

The proposed model is implemented with Pytorch that use 2D keypoints from HRNet detector [11], CPN Detector or 2D ground truth to analyze the performance. In this paper, the 2D pose detector was implemented based on AlphaPose [12] codebase while the 3D pose estimator followed the PoseFormer codebade [6]. 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 [13]. 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.

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 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 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 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 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.

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

C. 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 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.

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 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 4 were 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.

TABLE I

QUANTITATIVE COMPARISONS WITH STATE-OF-THE-ART ON HUMAN3.6M DATASET USING CPN DETECTOR UNDER PROTOCOL #1 AND PROTOCOL #2 FOR FULLY-SUPERVISED METHODS. THE BOLD NUMBER IS THE BEST PERFORMANCE IN EACH CASE, ⋆ DENOTES THAT THE 2D KEYPOINT DETECTION IS THE CASCADED PYRAMID NETWORK(CPN) WHILE ∗, † REFERS TO 3D NETWORK APPLY TRANSFORMER-BASED MODEL

Protocol # 1 - CPN	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Punch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [14]	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.* [15]	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.* [16]	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 [11]	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.* [3]	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.* [6]	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.* [15]	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.* [4] ⋆	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
Yang *et al.* [17]	**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.* [6]	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.* [16]	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 II

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.* [14]	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.* [15]	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.* [16] †	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 [11]	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.* [3]	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.* [18]	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.* [19]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32.3
Yang *et al.* [6]	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**

TABLE III

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*. [4] (f=81)	86.0	51.9	84.0
Lin *et al*. [8] (f=25)	83.6	51.4	79.8
Li *et al*. [16]	81.2	46.1	99.7
Chen *et al*. [19]	87.9	54.0	78.8
Yang *et al* [6] (f=9)	88.6	56.4	75.5
Our (f=9)	89.1	57.5	76.3

TABLE IV

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 [17]	45.7	37	3.02
Hourglass [20]	52.3	41.2	4.11
HRNet-W32 [11]	45.1	36.3	2.91
HRNet-W32+AM (our)	43.6	35.1	2.77
GroundTruth	28.6	24.5	0.98

TABLE V

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

TABLE VI

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

Hence, this paper decided to put SAM and TAM between the MSA and MLP.

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 [4], PoseFormer by PoseFormer et al. [6], 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 Posi-

TABLE VII
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

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

Method	Stride Frame	#Param (M)	MPJPE ↓
SimplePose *et al*. [4]	243 (default)	12.75M	47.5
SimplePose *et al*. [4]	81	12.70M	47.9
PoseFormer *et al*. [6]	81 (default)	9.59M	44.3
PoseFormer *et al*. [6]	27	9.60M	44.6
Our	9	9.85M	44.3
Our	27	9.86M	43.6
Our	81	9.86M	43.3

TABLE IX
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 ↓
SimplePose *et al*. [4]	0.25 (default)	12.70M	47.9
SimplePose *et al*. [4]	0.1	12.70M	47.5
PoseFormer *et al*. [6]	0.25 (default)	9.60M	44.3
PoseFormer *et al*. [6]	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

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.

V. CONCLUSION

This research shows the effect of the data augmentation on CNNs especially for occluded human keypoint, focusing on mosaic and mix-up for human proposals. Furthermore, our work demonstrates that not increasing the computation cost, the data augmentation utilized has a more considerable effect. Moreover, the mosaic and mix-up focused more on the essential feature map than the other element. The network will become more effective as a consequence, particularly for various computer vision-related tasks.

Besides, human pose estimation has several problems that need to be solved. First, the occluded joints were challenging to train and predict for the architecture. Second, human key points appear in the low-resolution images. The next issue is the sample has a crowd, which is usually difficult to identify where each participant's joint location. Last but not least, The lacking of data with partial body part appear with human posture. The proposed method tries to solve the first problem is also the most complex case compared to all of the issues. Hence, future research will try to focus on the remaining problem and also try to apply the technique to other state-of-the-art pose estimators.

ACKNOWLEDGEMENT

This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE) (2021RIS-003)

REFERENCES

[1] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *CoRR*, vol. abs/1312.4659, 2013. [Online]. Available: http://arxiv.org/abs/1312.4659
[2] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016. [Online]. Available: http://arxiv.org/abs/1603.06937
[3] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019.
[4] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, and J. Wang, "Human pose estimation using global and local normalization," 2017.
[5] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision (ACCV)*, 11 2012, pp. 31–44.
[6] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," 2017.
[7] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern discovery," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 48–53, Jan 2010.
[8] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *CoRR*, vol. abs/1904.05005, 2019. [Online]. Available: http://arxiv.org/abs/1904.05005
[9] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016.
[10] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," 2017.
[11] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," 2019.
[12] C. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation + matching," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5759–5767.
[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.
[14] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected u-nets for efficient landmark localization," 2018.
[15] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," 2016.

[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," 2016.
[17] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," 2017.
[18] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312
[19] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," *Lecture Notes in Computer Science*, p. 717–732, 2016.
[20] Z. Hussain, M. Sheng, and W. E. Zhang, "Different approaches for human activity recognition: A survey," 2019.