PoseGate-Former: Transformer Encoder with Trainable Gate for 3D Human Pose Estimation using Weakly Supervised Learning

Shannan Guan[1], Haiyan Lu[1], and Linchao Zhu[1] and Gengfa Fang[2]

[1] Australia Artificial Intelligence Institute
[2] Global Big Data Technologies Centre
University of Technology Sydney, Australia, AU

Abstract. Weakly supervised learning for 3D human pose estimation can learn a real human structure, but it generally has lower accuracy on reconstructing 3D poses. In this work, we present a 3D pose estimation model using a Transformer encoder based architecture with a trainable gate, **PoseGate-Former**. The model is trained using individual images from a weakly supervised learning approach. It can reduce possibility of overfitting on some action categories due to the addition of a trainable gate to the Transformer encoder. We evaluated this model on two benchmark datasets: *Human3.6M* and *HumanEva-I*. The experimental results show that this model can obtain substantially better accuracy in all action categories of 3D human poses in the datasets compared with some fully-supervised 3D pose estimation approaches.

Keywords: 3D Pose estimation· Transformer · Weakly-supervised learning

1 Introduction

3D human pose estimation from individual monocular images aims to predict the (x,y,z) coordinates of each key joint of a human body in the camera coordinate system. It is a challenging problem in the computer vision area as the relationship between 2D and 3D human poses can be one-to-many. There are two main approaches to estimate 3D human pose from monocular images: supervised approach and a weakly-supervised approach [4–6, 8, 11, 12]. Although it can learn a real human structure, the weakly-supervised approach generally has a lower accuracy in predicting the 3D joint coordinates. Therefore, it is highly desirable to develop a new method to improve the accuracy of 3D human pose estimation from the weakly supervised learning approach.

With the inspiration of the success of Transformer architecture in Computer Vision, we proposed a new 3D pose estimation models using a modified transformer encoder with a trainable gate, referred to as **PoseGate-Former**, and this model is trained by using a weakly supervised learning approach. Compared with fully connected neural networks, self-attention mechanism in a transformer architecture could learn the relations among human key joints and improve the

accuracy of 3D human pose estimation using a weakly supervised learning approach.

We evaluate our model on two benchmark datasets: 1) Human3.6M [2], and 2) HumanEva-I [10]. It has been observed that the **PoseGate-Former** can improve the estimation accuracy significantly in all action categories and improve the performance in a few specific action categories. It can reduce 30% in average MPJPE compared with RepNet [11], and can outperform most of supervised learning approaches.

Our contributions are twofold: 1) Introduced the self-attention architecture of a Transformer in 3D human pose estimation with significantly improved performance by using a weakly supervised training approach and 2) Proposed the PoseGate-Former by adding a trainable gate to the self-attention architecture of a Transformer to reduce the possibility of overfitting on some specific action categories, evidenced by our experimental results on Human3.6M and HumanEva-I.

2 Related Work

Two approaches in the literature are relevant to this study: One is 2D to 3D human pose conversion approaches based on individual images using a weakly supervised learning. For example, Wandt et al. [11] proposed a weakly supervised adversarial learning structure to lift human pose from 2D to 3D. The other one is the Transformer learning architecture. It is promising to use a Transformer architecture in the Computer Vision area. For examples, work in [4, 12] use a Transformer encoder architecture to map 2D human poses to 3D poses by using a sequence of 2D key joints extracted from videos.

3 Method

In this section, we first present our new design of a 3D pose estimator, the PoseGate-Former, then present the weakly supervised learning structure and lastly explain the learning procedure.

3.1 PoseGate-Former architecture for 3D Pose Estimation

As shown in Fig1 (a), the PoseGate-Former includes a trainable gate module, there are two branches feed to the multi-head self-attention module: The left branch is used for providing matrices $Q = (q_1, q_2, \ldots, q_h)$, $K = (k_1, k_2, \ldots, k_h)$, and $V = (v_1, v_2, \ldots, v_h)$, where q_i,k_i, and v_i is the i_{th} element in the matrices Q,K and V, respectively, $i = 1, 2, \ldots, h$. The right branch is split into two paths, one path outputs a h dimensional gate vector $G = (g_1, g_2, \ldots, g_h)$, and the other path outputs a h dimensional bias vector $B = (b_1, b_2, \ldots, b_h)$. In each path, two fully connected layers of 100 neurons with sigmoid activation function are used. The sigmoid activation function is used to limit the output value ranging

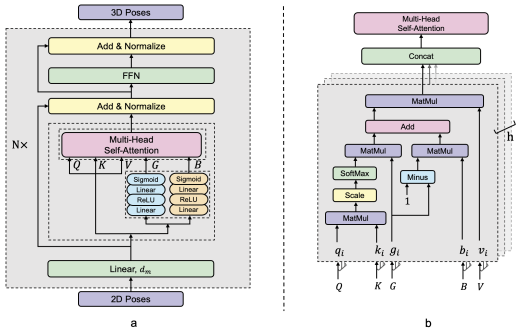

Fig. 1. Part (a) shows the structure of PoseGate-Former which is the 3D pose estimator. Part (b) illustrates the calculation procedure of a single head and the logic of how h heads of self-attention scores are concatenated to multi-head self-attention scores.

from zero to one. From Fig1 (b), the self-attention score of the head i can be calculated by:

$$\text{head}_{\ i} = \left(\text{softmax}\left(\frac{q_i k_i^T}{\sqrt{d_k}}\right) g_i + (1 - b_i)\, g_i\right) v_i \tag{1}$$

where g_i is a gate value from gate vector G and b_i is a bias value from bias vector B. Then, the multi-head self-attention score can be obtained by concatenating the scores of all heads: $\text{MultiHead}\ (Q, K, V, G, B) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)$. Based on Eq.1, the multi-head self-attention can be expressed as:

$$\begin{aligned}\text{Self-Attention}(Q, K, V, G, B) = (&\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{diag}(G) \\ &+ \text{diag}(I - B) \cdot \text{diag}(G))V\end{aligned} \tag{2}$$

where I is a h dimensional vector which only consists of one, and *diag* is a diagonal matrix. This structure can output corresponding gate and bias values based on different poses and effectively reduce the possibility of overfitting in some specific action categories by correcting the multi-head self-attention scores.

3.2 Weakly Supervised Learning Structure

The weakly supervised learning structure for training the PoseGate-Former is developed based on the training structure in [11], and shown in Fig2. In this structure, a 2D pose is fed into the PostGate-Former to generate an estimated 3D pose. Meanwhile, this 2D pose is also fed into the camera module which outputs a projection matrix M for simulating the camera projection. Then, 3D pose and the generated projection matrix M are fed into a projection module for projecting the corresponding 2D pose. During the training process, a critic module will judge whether the generated 3D pose corresponds to a real human shape.

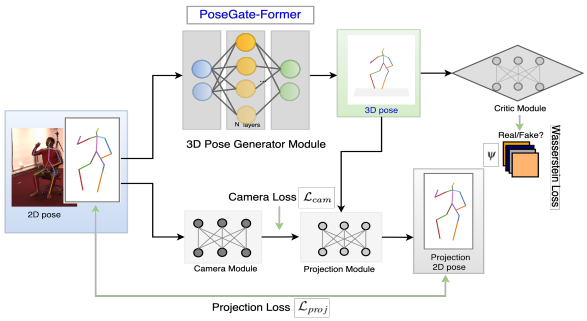

Fig. 2. The weakly supervised adversarial learning structure used in this work, which consists of four modules: 1) a 3D pose generator module, 2) a camera estimation module, 3) a projection module, and 4) a critic module.

3D Pose Generator Module. This module implements the PoseGate-Former. It's input is the extracted 2D key joints expressed as a matrix W, where $W \in \mathbb{R}^{2\times n}$ and its output is a $3n$ dimensional vector that consists of (x,y,z)-coordinates of each key joint and is reshaped to X. In following expression, $X \in \mathbb{R}^{3\times n}$ donates the 3D pose and the (x,y,z)-coordinates of each key joint are recorded in n columns that correspond to the inputs of this module, repectively.

Critic Module. This module is used to ensure that the generated 3D pose corresponds to a real human shape. The estimated 3D pose from the 3D pose generator will be transformed by a kinematic chain space (KCS) matrix [11] and fed into a fully connected neural networks to output a single critic value for calculating the Wasserstein loss ψ through a Wasserstein loss function [11].

Camera Module. In most instances, we do not know camera parameters to project 3D poses, therefore, we need to build a camera module to regress camera parameters to project 3D poses. The camera module regresses a vector with six parameters, and the vector is reshaped to the projection matrix $M \in \mathbb{R}^{2\times 3}$.

Projection Module. This module is used to transfer the output X from 3D pose generator network to a 2D pose matrix W' by multiplying the 3D pose matrix X with the projection matrix M from camera estimation network: $W' = MX$.

3.3 Training Procedure

In order to train the PoseGate-Former, we apply three losses to guide the training from the weakly supervised learning approach: 1) Wasserstein Loss ψ in the last layer of critic network, 2) Camera Loss $\mathcal{L}_{cam}$ to calculate the camera loss in the camera network, and 3) Projection Loss $\mathcal{L}_{proj}$ to minimize the errors between the ground-truth and the estimated poses. In the training procedure, we implemented the Improved Wasserstein GAN training method [1]. We group different modules into two models: 1) adversarial model and 2) discriminator model and train them separately. The adversarial model contains a complete

learning structure as shown in Fig2, but the critic module only implements the feed forward inference propagation without training the parameters. The discriminator model consists of the 3D pose generator module and the critic module, and only the critic module will be trained.

4 Model Evaluation and Discussion

There are two main evaluation protocols for evaluating the proposed methods, both of them use the mean per joint positioning error (MPJPE), which calculates the average Euclidean distance between the estimated joint and the corresponding ground truth joint coordinates. Protocol-I directly calculates the MPJPE. Protocol-II applies a rigid alignment between the ground truth and the estimated poses and calculates the P-MPJPE.

4.1 Quantitative Evaluation on Human3.6M

Human3.6M is the largest public 3D human pose estimation dataset. This dataset contains 15 categories of daily activities of 7 professional subjects. In this work, we used 5 subjects (1, 5, 6, 7, 8) for training and 2 subjects (9, 11) for evaluating. Table 1 shows the evaluation results on Human3.6M dataset using Protocol-I.

Table 1. Comparisons of MPJPE error from PoseGate-Former along with other state-of-the-art 3D post estimation methods. The column WS indicates whether this approach used a weakly-supervised method. The best are shown in bold, second-best are underlined.

Protocol-I	**WS**	**Direct.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Purch.**	**Sit**	**SitD**	**Smoke**	**Wait**	**Walk**	**WalkD**	**WalkT**	**Avg.**
Park et al. [7]		100.3	116.2	90.0	116.5	115.3	149.5	117.6	106.9	137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3
Zhou et al. [14]		91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Luo et al. [5]		68.4	77.3	70.2	71.4	75.1	86.5	69.0	76.7	88.2	103.4	73.8	72.1	83.9	58.1	65.4	76.0
Pavlakos et al. [8]		67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Zhou et al. [13]		54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.2	66.1	63.2	51.4	55.3	64.9
Martinez et al. [6]		53.3	60.8	62.9	62.7	86.4	82.4	57.8	58.7	81.9	99.8	69.1	63.9	50.9	67.1	54.8	67.5
Wandt et al. [11]	✓	50.0	53.5	44.7	51.6	49.0	58.7	48.8	51.3	51.1	66.0	46.6	50.6	42.5	38.8	60.4	50.9
PoseGate-Former (Ours)	✓	32.0	**34.3**	**26.9**	**34.6**	**37.8**	**35.7**	**27.8**	38.2	**34.9**	**38.7**	**31.4**	**39.8**	**38.5**	**37.0**	**40.2**	**35.2**

It can be seen from Table 1 that the proposed PoseGate-Former shows a significant improvement compared with other benchmark methods. For each action category, the PoseGate-Former is able to mitigate the overfitting problem and the errors in each category are rapidly reduced or remain the same. The average error reduced is 30% compared with the one from RepNet [11]. Fig.3 shows the comparisons of reconstructed 3D poses by our PostGate-Former and the ground truth 3D poses in the validation dataset in Human3.6M. It can be seen from Fig.3 that all the poses were well reconstructed, even complex poses, such as Sitting on the ground, Phoning, and Crossing legs.

4.2 Quantitative Evaluation on HumanEVA-I

Compared with Human3.6M, HumanEVA-I is a smaller dataset which contains three action categories (Walk, Jog, Box) performed by subjects (S1, S2, S3).

Table 2 shows the comparison results of P-MPJPE between our structure and other few state-of-the-art approaches on HumanEva-I. It can be seen that our PostGate-Former achieved promising performance. Compared with individual images based fully supervised approaches, our PoseGate-Former has a better performance across all action categories, including complex action categories (e.g. Walk S3 subject, Box action category). Compared with video based approach [4], our PoseGate-Former also achieved a comparable performance.

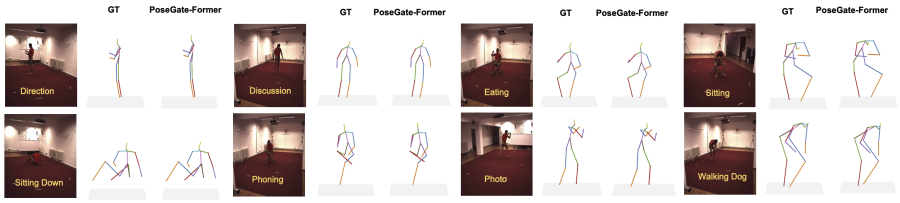

Fig. 3. Visualization examples of 3D pose reconstruction for some action categories from the validation dataset of Human3.6M. The GT columns show the ground truth poses and PoseGate-Former columns show the reconstructed 3D poses.

Table 2. Quantitative results for reconstructing 3D pose of HumanEva-I dataset following *Protocol-II*.Video means the approach is video based, and the best in bold, second-best underlined. Our results show P-MPJPE by using ground truth 2D labels.

HumanEVA-I	Walk			Jog			Box			Avg.
	S1	S2	S3	S1	S2	S3	S1	S2	S3	
Martinez et al. [6]	19.7	17.4	46.8	26.9	18.2	18.6	-	-	-	-
Pavlakos et al. [8]	22.3	19.5	29.7	28.9	21.9	23.8	-	-	-	-
Lee et al. [3]	18.6	19.9	30.5	25.7	16.8	17.7	42.8	48.1	53.4	30.3
Pavllo et al. [9]	13.9	10.2	46.6	20.9	13.1	13.8	23.8	33.7	32	23.1
Li et al. [4] (Video)	**9.7**	**7.6**	15.8	12.3	**9.4**	11.2	14.8	**12.9**	16.5	**12.2**
PoseGate-Former (Ours)	13.0	9.9	**15.7**	**11.9**	12.1	**10.23**	**12.4**	13.4	**12.1**	12.3

4.3 Ablation Study

To validate the contribution of key components of PoseGate-Former, e.g., the self-attention layer and the trainable gate, and the impact of hyperparameters on performance, we carried out an ablation study on Human3.6M dataset. This study is to verify the contributions made by the self-attention layer and the trainable gate to the performance of PoseGate-Former. In this study, we set the dimension of the Transformer architecture d_m to 256, and evaluate the contributions in each action category based on MPJPE. We implemented the structure/model under three conditions: 1) We fixed all self-attention scores to $1/n(n = 16)$. Because we take 16 key joints in Human3.6M and the sum of self-attention scores is one, thus the average value is 1/16; 2) We use a naive Transformer self-attention architecture; and 3) We only used one trainable value in the gate and used a constant value $1/n(n = 16)$ as the bias.

It can be seen from Table 3 that the attention scores have significant impact on the performance of our pose generator, PostGate-Former. Compared with the naive Transformer model, a fixed self-attention score leads MPJPE to increase

by 15%. The PoseGate-Former with a Fixed-bias column shows that one trainable value in the gate can enhance the performance by 9% compared with the naive Transformer model. It shows the best results in some action categories, such as Direction, Discussion, and Sitting. However, the overall performance under PoseGate-Former with a fixed bias is worse than PoseGate-Former due to overfitting in some specific action categories.

Table 3. Ablation study on different self-attention layers in the Transformer architecture. The results show MPJPE which are implemented on Human3.6M using *Protocol-I* with the ground-truth 2D poses as the inputs. Fix-attn is fixing all self-attention scores, Fix-bias is using one trainable value as gate and use a constant value as bias.

Ablation study 1	**Direct.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Purch.**	**Sit**	**SitD**	**Smoke**	**Wait**	**Walk**	**WalkD**	**WalkT**	**Avg. (↓)**
Transformer (Fix-attn)	42.0	42.1	41.3	54.2	46.6	51.1	43.8	49.3	46.0	48.6	53.5	57.7	50.2	55.7	53.1	48.7
Naive Transformer	31.5	35.5	33.3	41.7	39.5	43.5	33.6	37.8	37.8	66.6	35.2	47.5	46.8	41.7	50.0	41.5
PoseGate-Former (Fix-bias)	28.9	33.5	30.9	37.3	40.4	38.5	33.7	37.9	34.5	43.5	33.9	45.5	41.5	46.6	41.6	37.9
PoseGate-Former	32.0	34.3	26.9	34.6	37.8	35.7	27.8	38.2	34.9	38.7	31.4	39.8	38.5	37.0	40.2	35.2

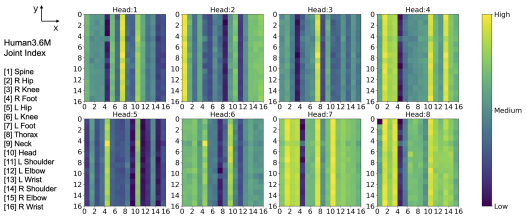

Fig. 4. Visualization of multi-head self-attentions in PoseGate-Former, the x-axis indicates input queries and y-axis show the predicted outputs. Yellow color indicates a stronger attention.

4.4 Self-Attention Visualization.

To illustrate the multi-head self-attention mechanism, we visualized the self-attention scores of PoseGate-Former. As shown in Fig 4, we can find that the Head 1 focuses on Thorax joint, Head 2 focuses Spine joint. Head 4 builds the connection among joints (11, 12, 13) and (14, 15, 16) which are grouped as left arm and right arm. For Head 8, it connects joints (2, 3, 4) which belong to right leg. These attention maps show that the PoseGate-Former successfully finds the relationship between key joints, and these relationships are hard to learn by fully-connect neural networks. This could explain why a Transformer architecture can significantly improve the performance of 3D pose estimation model.

5 Conclusions

In this work, we develop a Transformer based PoseGate-Former to lift 2D poses to 3D domain by using a weakly supervised learning approach. We found that the multi-head self-attention architecture in Transformer can easily learn the relationship among human key joints, which can significantly improve the performance of 3D human pose estimator. More importantly, our trainable gate

mechanism can effectively reduce the possibility of overfitting in some specific action categories compared with the naive Transformer architecture and further improve the performance of PoseGate-Former.

References

1. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. CoRR **abs/1704.00028 (2017)**, http://arxiv.org/abs/1704.00028
2. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2014). https://doi.org/10.1109/TPAMI.2013.248
3. Lee, K., Lee, I., Lee, S.: Propagating lstm: 3d pose estimation based on joint interdependency. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
4. Li, W., Liu, H., Ding, R., Liu, M., Wang, P.: Lifting transformer for 3d human pose estimation in video. CoRR abs/2103.14304 (2021), https://arxiv.org/abs/2103.14304
5. Luo, C., Chu, X., Yuille, A.L.: Orinet: A fully convolutional network for 3d human pose estimation. vol. abs/1811.04989 (2018), http://arxiv.org/abs/1811.04989
6. Martinez,J.,Hossain,R.,Romero,J.,Little,J.J.:Asimpleyeteffectivebaselinefor 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
7. Park, S., Hwang, J., Kwak, N.: 3d human pose estimation using convolutional neural networks with 2d pose information. In: Hua, G., J egou, H. (eds.) Computer Vision –ECCV 2016 Workshops. pp. 156–169. Springer International Publishing, Cham (2016)
8. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
9. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 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 (CVPR) (June 2019)
10. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision **87**, 4–27 (2009). https://doi.org/10.1007/s11263-009-0273-6
11. Wandt, B., Rosenhahn, B.: Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation (June 2019)
12. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. CoRR **abs/2103.10455 (2021)**, https://arxiv.org/abs/2103.10455
13. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: A weakly-supervised approach. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
14. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: Hua, G., J egou, H. (eds.) Computer Vision – ECCV 2016 Workshops. pp. 186–201. Springer International Publishing, Cham (2016)