IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 May 2022, accepted 11 June 2022, date of publication 21 June 2022, date of current version 29 June 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3185058

Learning 3D Skeletal Representation From Transformer for Action Recognition

JUNUK CHA[1], MUHAMMAD SAQLAIN[1], DONGUK KIM[1], SEUNGEUN LEE[2], SEONGYEONG LEE[2], AND SEUNGRYUL BAEK[1]

[1]School of Artificial Intelligence, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea
[2]School of Computer Science, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea

Corresponding author: Seungryul Baek (srbaek@unist.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) (Development of Human Image Synthesis and Discrimination Technology Below the Perceptual Threshold) under Grant 2021-0-01778; in part by the Artificial Intelligence Graduate School Program, Ulsan National Institute of Science and Technology (UNIST) under Grant 2020-0-01336; in part by the Development of 5G Based Low Latency Device–Edge Cloud Interaction Technology under Grant 2020–0–00537; in part by the Artificial Intelligence Innovation Hub under Grant 2021-0-02068; in part by the Settlement Research Fund of the UNIST under Grant 1.200099.01; and in part by the Grant from the Research and Development Program of the Korea Railroad Research Institute, Republic of Korea.

ABSTRACT Skeleton-based human action recognition has attracted significant interest due to its simplicity and good accuracy. Diverse end-to-end trainable frameworks based on skeletal representation have been proposed so far to map the representation to human action classes better. Most skeleton-based human action recognition approaches are based on the skeletons, which are heuristically pre-defined by the commercial sensors. Nevertheless, it is not confirmed whether the sensor-captured skeletons are the best representation of human bodies for the action recognition task, while in general, the dedicated representation is required for achieving the successful performance on subsequent tasks such as action recognition. In this paper, we try to deal with the issue by explicitly learning the skeletal representation in the context of the human action recognition task. We start our investigation by reconstructing 3D meshes of the human bodies from RGB videos. Then we involve the transformer architecture to sample the most informative skeletal representation from reconstructed 3D meshes, considering the inner and inter structural relationship of 3D meshes and sensor-captured skeletons. Experimental results on challenging human action recognition benchmarks (i.e., SYSU and UTD-MHAD datasets) have shown the superiority of our skeletal representation compared to the sensor-captured skeletons for the action recognition task.

INDEX TERMS 3D representation, action recognition, human mesh, transformer.

I. INTRODUCTION

Recognizing the temporal actions of human bodies has been an essential task in the field of computer vision. Several deep learning architectures [1]–[9] have been proposed so far to capture the human actions properly. Popular methods for the human body actions are based on RGB-D representation [1]–[4] and skeleton representation [5]–[9]. RGB-D video is the most trivial input that requires only RGB-D cameras to collect; however, it requires a large amount of memory capacity. In the aspect of efficiency, skeletal representation is the most promising, as it captures key component of human bodies and is able to achieve comparable accuracy to the RGBD-based human action recognition. Despite its success, the sensor-captured skeletons are varied depending on the sensor types, and there might be room for improvement in its representation.

Recognizing the 3D skeleton has been widely studied for its practicality after it was initially commercialized by the Kinect sensor [10]. There has been much technical progress in the field of 3D human pose estimation using the deep learning approaches [11]–[13]. More recently, many

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

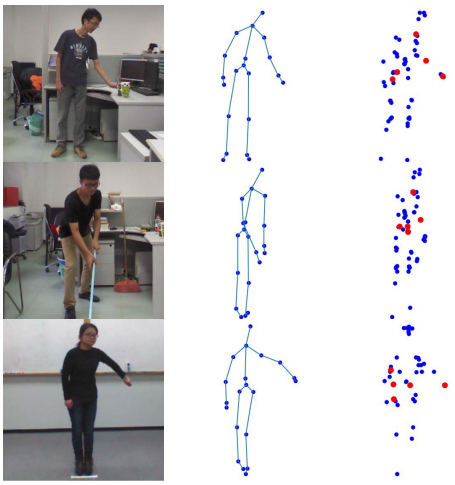

FIGURE 1. Examples of 3D skeletal representation obtained from the Kinect sensor and our algorithm: (a) original RGB images, (b) 3D skeletons obtained from the Kinect sensor, and (c) 3D skeletons obtained from our algorithm. For our representation, five points with high importance are highlighted in red color.

methods appeared to reconstruct both 3D poses and shapes of the human bodies from single RGB images. Most of them [14]–[16] are reconstructing 3D human meshes based on SMPL [17]. 3D meshes represent human bodies as thousands of 3D vertices, thereby capturing more detailed movements of the human bodies compared to coarse skeletons. 3D meshes have the potential to provide better action recognition capability than the sensor-captured skeletons.

In this work, we are motivated to explore to obtain the proper skeletal representation from human mesh for recognizing human actions. We developed our framework based on two steps: we first reconstructed 3D meshes of human bodies from RGB videos and then learned to sample the most informative skeletal representation, as shown in Figure 1, for action recognition using the transformer architecture [11]. We tried to reveal two key propositions via the proposed framework: 1) how much the 3D meshes could help the action recognition task and 2) which is the most effective configuration of skeletal representation for better action recognition.

At a more detailed level, reconstructing 3D meshes of human bodies from single RGB images is performed using the ExPose [16] algorithm. First, it reconstructs the 3D meshes via the SMPL-X model [15], which has $10,475$ vertices and $20,908$ triangular faces for the human's faces, bodies, and hands. Then, we involved the transformer architecture [11] in sampling the most informative skeletal representation for action recognition. Our transformer is mainly trained by involving three supervisions: considering the inner and inter relationship among 3D mesh vertices sequences input and 3D sensor-captured skeletons sequences input, and enforcing 1) our 3D sampled skeletons output close to the 3D mesh vertices input, 2) our 3D vertices output close to the 3D mesh vertices input, and 3) the excellent accuracy in the action recognition task. By the method, we showed the superiority of our learned skeletal representation compared to the sensor-captured skeletons on two challenging human action recognition benchmarks: SYSU [18] and UTD-MHAD [19] datasets. Our contributions could be summarized as follows:

- We propose to investigate a method that is able to learn the skeletal representation of human bodies for the action recognition task.
- We propose an effective skeletal representation sampling scheme by first reconstructing the 3D meshes of human bodies from RGB videos and then generating the most informative skeletal representation among them.
- We conduct experimental analysis on our skeletal representation learning method and then show that our method is able to produce the skeletal representation that has superior accuracy on SYSU and UTD-HMAD datasets compared to sensor-captured skeletons.

II. RELATED WORKS

In this section, we review 3D human mesh reconstruction algorithms that are able to capture human body motions. Then, we further review action recognition algorithms that are based on different modalities: RGB, depth, and 3D skeletons.

A. 3D HUMAN MESH ESTIMATION

Recently, there have been approaches for estimating both poses and shapes of humans in RGB images or videos. These methods provide the J-regression matrix that is able to obtain the 3D skeletons from the 3D mesh vertices. These 3D skeletons include the pre-defined skeletons such as wrists, elbows, ankles, knees, neck, head, etc.. There has been a method [20] that uses the advantage of both optimization-based and regression-based methods for estimating 3D mesh. Kocabas et al. [21] proposed a framework that incorporates the temporal dynamics of the human body and shape. Lin et al. [22] proposed the end-to-end human pose and mesh reconstruction with transformer [11], which is a non-parametric body mesh reconstruction method. 3D skeletons from these methods may not be the best tool for action recognition because these methods use SMPL [17] body model and this model can not express the pixel-aligned human hand meshes which provide important information to recognize action. Thus, we propose a framework that samples skeletal representation from SMPL-X [15] 3D meshes, which can express the human body, hand and face, for action recognition.

B. HUMAN ACTION RECOGNITION

Action recognition is a critical computer vision field where a lot of research is being done because it can be beneficial in real life. Therefore, many network models using various cues have been proposed: Zhang et al. [23] proposed semantics-guided neural networks for skeleton-based human action recognition. They introduced the semantics of skeletons and frames. It is more helpful to understand the relationship between skeletons and recognize human

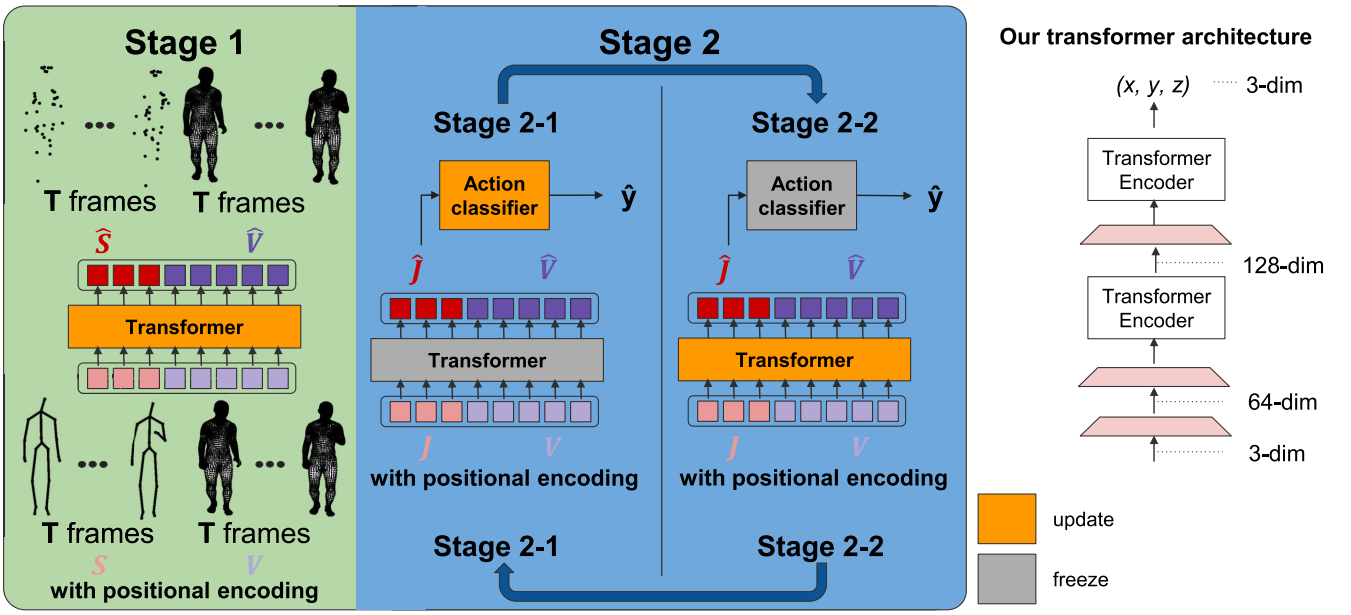

FIGURE 2. The summary for the overall training method: orange and gray blocks mean the model which needs to be trained and the frozen model whose parameters are fixed during training, respectively. The skeletons sequences input S and 3D mesh vertices sequences input V obtained from the ExPose are feed to the transformer with positional encoding. Then, sampled skeletons sequences output $\hat{S}$ and estimated vertices sequences output $\hat{V}$ are obtained from the transformer architecture. The action classifier (SGN) needs sampled skeletons sequences $\hat{S}$ as input. Then, it outputs the prediction of the action label $\hat{y}$. At Stage 1, the transformer is initially trained. At Stage 2, alternating training is performed and this is divided into two stages: 2-1 and 2-2. At Stage 2-1, the transformer is frozen and the action classifier is trained. At Stage 2-2, the transformer is trained and the action classifier is frozen. By this, the action classifier is trained to use the learned skeletal representation as input, and the transformer is trained to sample better skeletal representation dedicated to the action recognition task.

action. Das et al. [24] proposed a 3D sensor-captured skeletons and RGB-based network model with spatial embedding and attention. This model uses two modalities to learn better spatio-temporal features for action recognition. Zhang et al. [25] proposed a view adaptive neural network model to handle the challenge of large view variations in human actions. Zhang et al. [26] proposed the VA-fusion method that has a view adaptation sub-network. It selects the suitable observation view for action recognition. There are two network models, VA-RNN and VA-CNN, and the output features from the two network models are fused to predict action labels. Zhang et al. [27] proposed an element-wise-attention GRU network model. The input data is multiplied by the attention value and passed to the GRU model. This simple method can be applied to RNN and LSTM. Islam et al. [28] proposed a hierarchical multi-modal attention-based human activity recognition algorithm. It uses N modalities for action recognition. This model extracts the uni-modal spatio-temporal feature and uni-modal self-attention in each modality. Then, after fusing the uni-modal features, it computes the multi-modal features and multi-modal multi-head self-attention for action recognition. It uses RGB images, skeletons and inertial, *etc*. Liu et al. [29] proposed a multi-modal CNN-based action recognition network model. CNN module predicts a body heatmap and pose. The late fusion scheme improved the accuracy by using 3D sensor-captured skeletons and estimated heatmap. Wang et al. [30] proposed a multi-stream interaction network. It consisted of a human skeleton module, an object module, and the interaction between human and object modules. This network learns the relationship between humans and objects to recognize human action in a better way. Ke et al. [31] proposed to use the discriminator to learn latent long-term global information and local action information for action recognition. McNally et al. [32] proposed a spatio-temporal activation model using RGB videos for action recognition. Zhao et al. [33] proposed the probabilistic model, which is hierarchical. Bayesian framework improved the ability to detect intra-class variations in the spatial and temporal extent of actions. Weiyao et al. [34] proposed the multi-modal action recognition model, which consists of bi-linear pooling and an attention network. This model used RGB videos and skeletons.

Recently, there appears action recognition pipelines that exploit the transformer architecture [11]; while most approaches built their framework based on the skeleton representation obtained from the commercial sensors. We try to investigate more optimal skeletal representation for the action recognition in the aspect of the configurations and the number of skeletons.

III. METHOD

This paper aims to learn the proper skeletal representation for the human action recognition task. We will investigate the proper configurations and the number of the skeletons which are suitable for human action recognition. We have pre-stage

for reconstructing full-body 3D meshes from RGB videos involving the recently proposed 3D mesh reconstruction network [16]. We build our entire framework to have overall two distinct stages: involving the transformer architecture [11] to sample the most informative skeletal representation from 3D meshes for action recognition and involving the action classifier [23]. In the remainder of this section, we will explain the details of the positional encoding and how the individual stages work. Also, in Figure 2, we visualized our overall pipeline having a transformer-based sampling skeleton network and action classifier within it. We summarized our entire training process in Algorithm 1.

A. DETAILS ABOUT THE POSITIONAL ENCODING

In the transformer architecture [11], sequence order information is missing as the architecture does not have the recurrent layers which are able to encode the sequence feature. Thus, the transformer usually encodes such information using the positional encoding scheme. We also perform it to encode the skeletons and mesh vertices as the sequences having spatial information. In our preliminary experiment, we observed that without the positional encoding, the transformer training is not robust, thus we tried to involve it in our framework. We added ''spatial positional encoding'' to the skeletons and mesh vertices similar to previous transformer-based approaches [11], [22], [35]. Especially, we followed the method of [35] adding the learnable weight parameter $W \in \mathbb{R}^{(\#S+\#V)\times 64}$ to the skeletons and mesh vertices input embeddings, where $\#S$ is the number of skeletons and $\#V$ is the number of mesh vertices.

B. PRE-STAGE: EXTRACTING FULL-BODY 3D MESHES FROM RGB VIDEOS

To reconstruct the 3D meshes, we used SMPL-X [15] deformable 3D mesh model for human bodies and involved the ExPose [16] algorithm. SMPL-X full-body mesh has 10,475 vertices and 20,908 triangular faces. ExPose reconstructs the SMPL-X full-body 3D human mesh from a single RGB image. ExPose is pre-trained on a large 3D human mesh dataset, and it is one of the state-of-the-art methods in human pose and shape estimation. Please note that we do not fine-tune ExPose on SYSU and UTD-MHAD datasets. We reconstructed and saved the 3D human meshes from RGB images of the SYSU and UTD-MHAD datasets using ExPose [16] before starting the experiment.

C. STAGE 1: LEARNING TO SAMPLE THE SKELETAL REPRESENTATION

After reconstructing 3D meshes using ExPose [16], we sample the most informative skeletal representation for action recognition using the transformer. We encode the inner and inter relationship among 3D mesh vertices input V and 3D skeletons input S using the self-attention mechanism. From this mechanism, we can sample the skeletons $\hat{S}$ which retain the geometric information of 3D meshes and have implicit information of 3D skeletons S. The transformer needs 3D skeletons input S and 3D mesh vertices input V with positional encoding, and it outputs sampled skeletons sequences $\hat{S}$ and estimated vertices sequences $\hat{V}$. We proposed to define two losses L_{cf} and L_{rec} to train the transformer architecture at stage 1.

L_{cf} is the loss that is similar to the chamfer distance loss proposed in [36] that is proposed to make our sampled skeletons output $\hat{S}$ close to the 3D mesh vertices input V as follows:

$$L_{cf} = L_f(\hat{S}, V) + \beta L_m(\hat{S}, V) + \gamma L_b(\hat{S}, V) \tag{1}$$

where

$$L_f(\hat{S}, V) = \frac{1}{|\hat{S}|} \sum_{s\in\hat{S}} \min_{v\in V} ||s - v||_2^2 \tag{2}$$

$$L_m(\hat{S}, V) = \max_{s\in\hat{S}} \min_{v\in V} ||s - v||_2^2 \tag{3}$$

$$L_b(\hat{S}, V) = \frac{1}{|V|} \sum_{v\in V} \min_{s\in\hat{S}} ||v - s||_2^2. \tag{4}$$

In the average worst case, two losses L_f and L_m keep the sampled skeletons in $\hat{S}$ close to those in the 3D mesh vertices input V, respectively. On the other hand, L_b ensures that the sampled skeletons output $\hat{S}$ are well spread over 3D mesh vertices input V.

The second loss L_{rec} is proposed to minimize the absolute difference between our estimated 3D vertices output $\hat{V}$ and

Algorithm 1 The Summary of Our Entire Training Process

Input: For transformer's input, skeletons sequences S and input 3D mesh vertices sequences from ExPose V. For action classifier, sampled skeletons sequences $\hat{S}$.
Output: Sampled skeletons sequences $\hat{S}$ and estimated vertices sequences $\hat{V}$ from transformer. Prediction of action label $\hat{y}$ from action classifier.
Parameter: The number of epochs T_1, T_{2-1}, T_{2-2}, and T_3.
Target: Ground-truth action label y.

```
for t1 = 1, ..., T1 do
    For S, Ŝ, V, and V̂, calculate the loss from Eq. 6 and update parameters of transformer.
end for
for t2 = 1, 2, 3 do
    for t2-1 = 1, ..., T2-1 do
        For y and ŷ, calculate the loss from Eq. 7 and update parameters of the action classifier.
    end for
    for t2-2 = 1, ..., T2-2 do
        For S, Ŝ, V, V̂, y, and ŷ, calculate the loss from Eq. 8 and update parameters of transformer.
    end for
end for
for t3 = 1, ..., T3 do
    For y and ŷ, calculate the loss from Eq. 7 and update parameters of the action classifier.
end for
```

TABLE 1. Ablation study on SYSU and UTD-MHAD dataset using the different types of skeleton representation. The method in the first row shows the results obtained using the sensor-captured skeletons, while the second row shows the results obtained using our skeletal representation.

Method	SYSU		UTD-MHAD
	CS	SS	CS
sensor-captured skeleton + SGN	81.8	80.5	94.7
Our skeleton + SGN	**88.0**	**86.3**	**96.3**

the 3D mesh vertices input V as follows:

$$L_{rec} = ||V - \hat{V}||_1. \tag{5}$$

L_{rec} helps the our estimated 3D vertices output $\hat{V}$ does not lose the information of 3D mesh vertices input V and sampled 3D skeletons output $\hat{S}$ can be affected by self-attention from well-reconstructed estimated 3D vertices output $\hat{V}$ and it can retain the geometry information of 3D mesh vertices input V. Thus, sampled 3D skeletons output $\hat{S}$ can be sampled well from the 3D mesh vertices input V.

We used the following weighted sum of the above two-loss terms for training transformer at stage 1:

$$L_{s1} = \alpha L_{cf} + \delta L_{rec} \tag{6}$$

where $\alpha = 30$, $\beta = 1$, $\gamma = 1$ and $\delta = 10$.

D. STAGE 2: LEARNING TO RECOGNIZE ACTION AND TO SAMPLE SKELETAL REPRESENTATION TOGETHER

In stage 2, we involved the action recognition network (i.e., SGN [23]) to simultaneously train the action classifier using the newly obtained skeletal representation and further optimize the transformer using the supervision from the action classifier. We involved additional cross-entropy loss for the training action classifier (i.e., SGN) as follows:

$$L_{ce} = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \tag{7}$$

where C is the number of action labels, y_i is the ground-truth action label, and $\hat{y}$ is the probability of action prediction.

Stage 2 is divided into two stages: stage 2-1 and stage 2-2. During the stage 2-1, the action classifier is trained using the cross-entropy loss L_{ce}, while during the stage 2-2, the transformer is further trained using the loss as follows:

$$L_{s2} = \alpha L_{cf} + \delta L_{rec} + \epsilon L_{ce} \tag{8}$$

where $\alpha = 30$, $\delta = 10$, and $\epsilon = 0.01$. Using cross-entropy loss, the transformer can be trained to sample the most informative skeletal representation that is good for action recognition. We conducted an experiment on the effectiveness of L_{ce} for training a transformer. More experimental details can be found in the Experiments section.

IV. EXPERIMENTS

We have used SYSU [18] and UTD-MHAD [19] benchmarks and reconstructed 3D mesh of human bodies from each RGB image using ExPose [16]. Its vertices are used as input of transformer-based sampling skeleton network with sensor-captured 3D skeletons.

A. DATASET

1) SYSU HUMAN-OBJECT INTERACTION DATASET [18]

SYSU dataset contains 12 action classes and 40 different subjects. It has 480 videos and provides 20 skeletons with 3D coordinates for each video. We used two protocols, Cross Subject (CS) and Same Subject (SS), proposed in [18]. For Cross Subject, half of the subjects are used for training and the others for testing. For the Same Subject, half of the sequences for each subject are used for training and others for testing. We reported the average accuracy of 30-fold cross-validation.

2) UTD-MHAD [19]

UTD-MHAD contains 27 action classes and 8 different subjects. 8 subjects repeated each action four times. It has 861 videos and provides 20 skeletons with 3D coordinates for each video. It also has depth and inertial data modalities. We used Cross Subject for the evaluation proposed in [19]. The data for the subject numbers 1, 3, 5, 7 were used for training, and the data for the subject numbers 2, 4, 6, 8 were used for testing.

B. IMPLEMENTATION DETAILS

When 10, 475 mesh vertices are fed into the transformer as input at once, it causes an insufficient memory error. Our transformer is proposed to process the coarse mesh vertices to prevent this. We used an average pooling layer to pool 10, 475 mesh vertices to 431 mesh vertices. We selected the average pooling layer over max-pooling layer as it can keep the geometric information contained in the original 10, 475 mesh vertices. As a result, the number of mesh vertices for the transformer are 431.

1) DATA PROCESSING

We divided each action video into 20 clips equally. We generate a new sequence of 20 frames by randomly choosing one frame from each clip for training. For testing, we create 5 new sequences as we did in training and used the mean score to predict the action label. During training, data augmentation is used. For SYSU and UTD-MHAD datasets, the 3D mesh X, Y, Z angle rotation is used for data augmentation. We randomly select the three angle rotation degrees between [-30, 30] for X, Y, Z axes, respectively.

2) TRAINING

We used the PyTorch library with Titan GPU. We used the Adam optimizer [37] with a learning rate of 0.001 for the transformer and SGN. The batch size is set to 16 for all datasets. The Label smoothing [38] is used for action recognition, and smoothing factor 0.1 is used. We also set T_1, T_{2-1}, T_{2-2}, and T_3 to 100, 20, 20, and 120, respectively, in Algorithm 1.

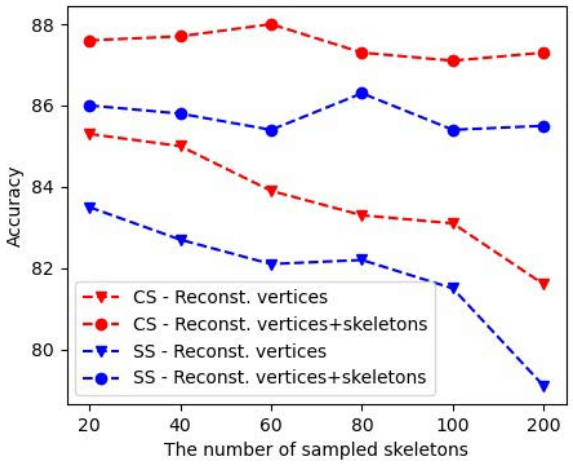

FIGURE 3. Ablation study on SYSU dataset using the different number of sampled skeletons and different types input of transformer: From this graph, we can see the overall action recognition accuracy is affected by the different number of skeletons, and we could obtain the best results using around 60 to 80 skeletons. The triangle marker denotes the method using only reconstructed 3D mesh vertices for transformer input. The circle marker denotes the method using both sensor-capture skeletons and reconstructed 3D mesh vertices for transformer input. Blue and red lines denote the same subject (SS) protocol and cross subject (CS) protocol, respectively.

C. ABLATION STUDY

We conduct ablative experiments for better understand our results. Especially, we designed and conducted two types of ablative experiments (for varying the number of skeletons and for varying the loss functions used) as follows:

1) THE DIFFERENT NUMBER OF SKELETONS AND DIFFERENT TYPES INPUT

In Table 1, Figure 3, and Figure 5, we experiment with the different number of sampled skeletons and different types of input on the SYSU and UTD-MHAD datasets. In Table 1, the skeletal representation obtained from our proposed transformer is better than the sensor-captured skeletons for action recognizer (SGN) input. In the SYSU in CS protocol, SYSU in SS protocol, and UTD-MHAD in CS protocol, our proposed method achieves the 6.2%, 5.8%, and 1.6% higher accuracy, respectively. As shown in Figure 3, the best number of sampled skeletons is 60 in CS protocol and 80 in SS protocol for the SYSU dataset. As shown in Figure 5, the best number of sampled skeletons is 40 or 80 for the UTD-MHAD dataset.

2) EFFECTIVENESS OF VARIOUS LOSSES FOR TRANSFORMER

We wanted to train the transformer to sample skeletons which are helpful for action recognition. To do that, we used L_{cf}, L_{rec}, and L_{ce}. L_{cf} makes sampled skeletons output $\hat{S}$ closer to 3D mesh vertices input V. L_{rec} makes estimated 3D vertices output $\hat{V}$ closer to 3D mesh vertices input V. L_{ce} makes transformer-based network to sample skeletons more helpful for action recognition. In Table 2, we observed that using all losses improves the accuracy.

TABLE 2. Effectiveness of L_{cf}, L_{rec}, and L_{ce} for training transformer to sample skeletons which are helpful for action recognition. We conducted the experiment on the SYSU dataset in Cross Subject (CS) protocol.

Methods	#Sampled skeletons	Accuracy
L_{rec}	60	86.5
L_{cf}	60	87.0
$L_{cf} + L_{rec}$	60	87.1
$L_{cf} + L_{rec} + L_{ce}$	60	**88.0**

D. VISUALIZATION OF ATTENTION

We visualize the self-attention between a specified skeleton and all other vertices in Figure 4. The action label in the first row is *drinking* and in the second row is *basketball shoot*. The brighter the color, the stronger the attention. In the first row and column (f), the skeleton on the left foot pays attention to all vertices except the right leg vertices. In the second row and column (f), the skeleton on the left knee pays attention to head vertices and arms vertices. Some skeletons attend to vertices close to them and some skeletons attend to vertices far from them for action recognition. Column (g) shows the output vertices (white) and sampled skeletons (red).

E. VISUALIZATION OF SMP

In Figure 6, we visualize the response of the spatial Max-Pooling layer in SGN [23]. Five green and red circles are top-5 skeletons selected by SMP in column (a) and column (b), respectively. Column (a) shows the important green skeletons for action recognition among blue skeletons captured by Kinect. Column (b) shows the important red skeletons for action recognition among blue skeletons obtained from the transformer. Column (c) shows the overlay on an original image with green and red circles in column (a) and column (b), respectively. The visualization shows the improvement in the correlation between learned skeleton joints and target actions compared to the correlation between conventional skeleton joints and actions: Especially, we observed that 1) more suitable skeleton locations are learned via the proposed method, thereby 2) the importance of skeleton joints are better captured and 3) action recognition accuracy is further improved accordingly. Five joints are chosen following the setting of [23], which is the suitable number out of 20 joints. In the first row and column (a), SGN considers the skeleton on the neck important for packing backpacks. However, in the first row and column (b), SGN considers skeletons on only hands important for *packing backpacks*. Intuitively, the hand skeletons are important to recognize packing backpacks. Likewise, in the second row and column (a), SGN considers the skeleton on the neck important for pouring. In the second row and column (b), SGN considers skeletons on the hands and the elbow important for *pouring*.

1) DISCUSSION

One limitation of our representation is: ours cannot learn the bone connection between the learned skeleton joints. Thus,

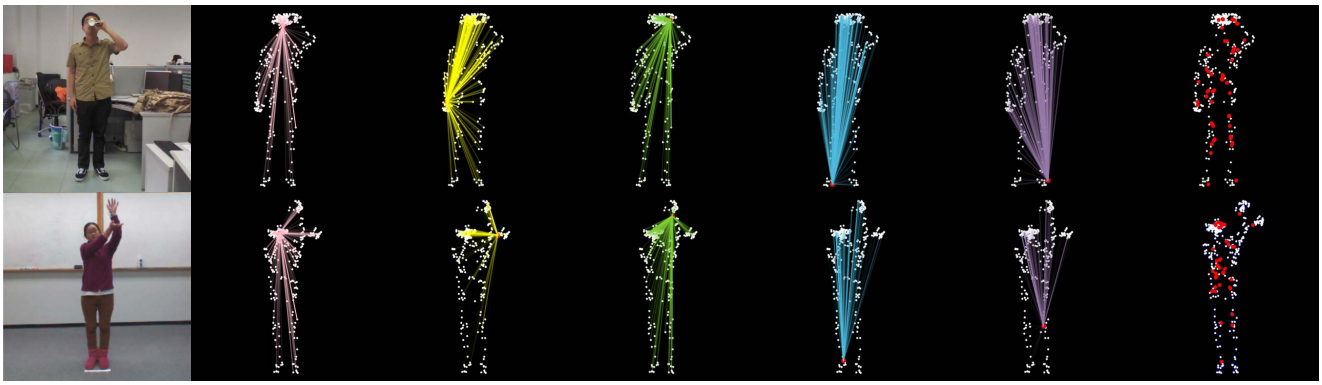

FIGURE 4. We visualize the self-attention between a specified skeleton and all other vertices. The brighter the color, the stronger the attention. The first row shows *drinking*, and the second row shows a *basketball shoot*. Column (a) shows original images. Column (b) shows self-attention of the skeleton on the head. Column (c) shows self-attention of the skeleton on the right arm. Column (d) shows self-attention of the skeleton on the left arm. Column (e) shows self-attention of the skeleton on the right leg. Column (f) shows self-attention of the skeleton of the left leg. Column (g) shows output vertices (white) and sampled skeletons (red).

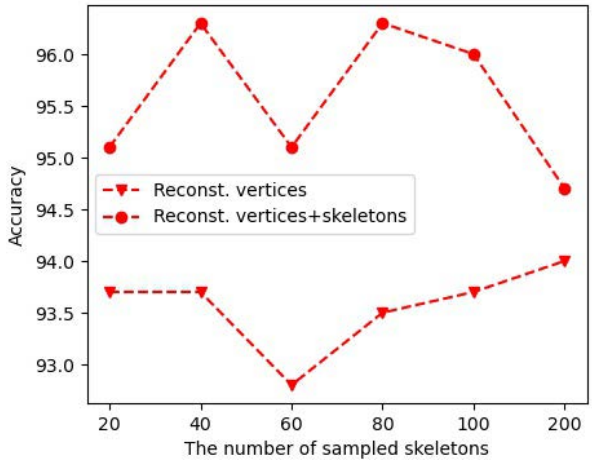

FIGURE 5. Ablation study on UTD-MHAD dataset using the different number of sampled skeletons and different types input of transformer: From the graph, we found that we can obtain the best action recognition accuracy using around 40 and 80 skeletons. The experiment is conducted using the Cross Subject (CS) protocol. The triangle marker denotes the method using only reconstructed 3D mesh vertices for transformer input. The circle marker denotes the method using both sensor-captured skeletons and reconstructed 3D mesh vertices for transformer input.

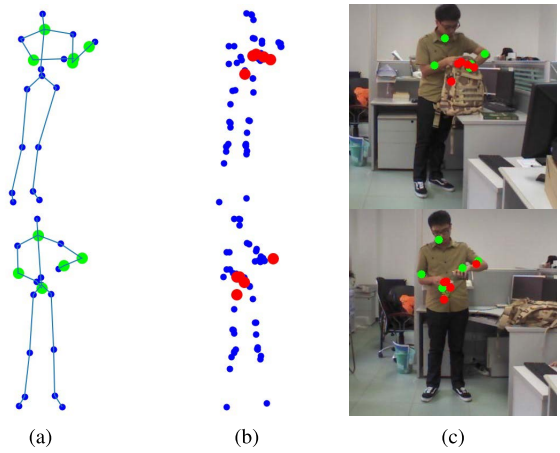

FIGURE 6. We visualize the responses of the spatial Max-Pooling layer in SGN. Column (a) is the result using the skeletons captured by Kinect with blue circles. Column (b) is the result using sampled skeletons obtained from the transformer with blue circles. Five green and red circles are top-5 skeletons selected by SMP among blue circles in column (a) and column (b), respectively. Column (c) shows top-5 skeletons with the original image. The first row and second row show *packing backpacks* and *pouring*, respectively.

the bone connection is missing in Figure 6(b) while there is the bone connection for conventional skeletons in Figure 6(a). We found that jointly learning both the bone connection and the skeleton joint locations is non-trivial due to the memory inefficiency: The bone connection exhibits N^2 complexity, if there is N joint locations. We think this challenge could be tackled in the future work. One thing to note is: recent action recognition pipelines such as [23] automatically learn to extract the correlation among skeleton joints without given spatial bone connections, which might be the crucial information for the action recognition. Thus, in our experiment, the spatial connection was not importantly tackled.

F. COMPARISON WITH STATE-OF-THE-ART METHODS

We analyzed our results in this sub-section. In Table 3 and Table 4, we enumerate the accuracy we obtain for both SYSU [18] and UTD-MHAD [19] datasets, respectively.

TABLE 3. Results for action recognition on SYSU dataset. * denotes the model uses parameters pre-trained on another large action dataset.

Method	Year	CS	SS
VA-LSTM [25]	2017	77.5	76.9
MSIN Human [30]	2021	82.0	80.3
SGN [23]	2020	83.0	81.6
LGN [31]	2019	83.3	-
EleAtt-GRU* [27]	2019	85.7	85.7
VA-fusion [26]	2019	86.7	86.2
Ours	-	**88.0**	**86.3**

In Table 3, we present the accuracy of our action classifier and other state-of-the-art methods on the SYSU dataset. There are two protocols, Cross Subject (CS) and Same Subject (SS). VA-LSTM [25], MSIN Human [30], SGN [23], LGN [31], EleAtt-GRU [27], and VA-fusion [26] require the skeletons input like our method. EleAtt-GRU uses weights pre-trained on another large action dataset, while our method

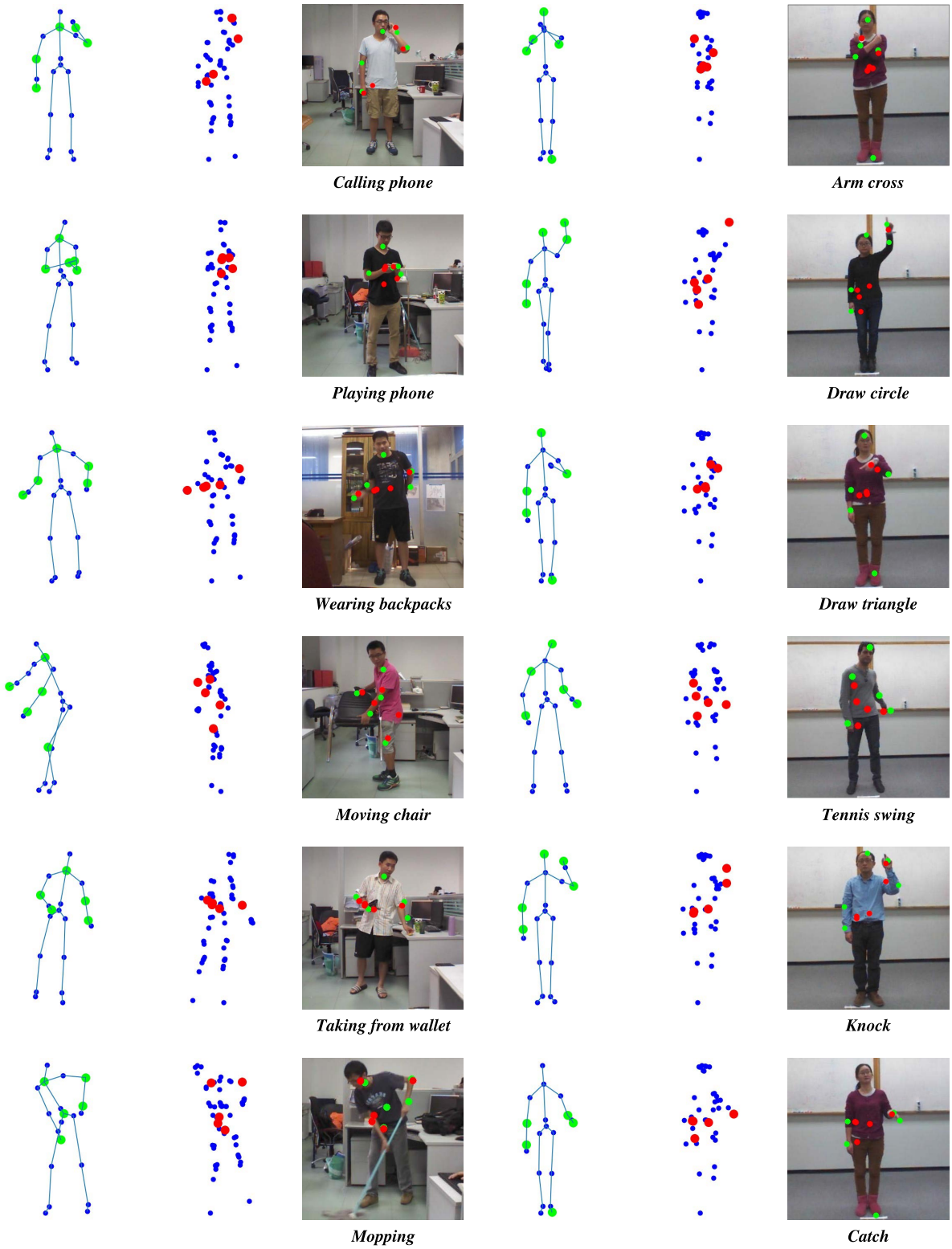

FIGURE 7. Visualization for responses of the spatial Max-Pooling layer in SGN. The 1st and 4th columns show results using the skeletons captured by Kinect. The 2nd and 5th columns show results using our sampled skeletons from the transformer. Five green and red circles are top-5 skeletons selected by SMP among green circles in the 1st and 4th columns, and the 2nd and 5th columns, respectively. The 3rd and 6th columns show top-5 skeletons visualized with the original image.

outperforms it and other state-of-the-art methods with our learned skeletal representation; while not using such a strong prior.

In Table 4, we present the accuracy of our human body action classifier and other state-of-the-art methods on the UTD-MHAD dataset. There is a protocol, Cross Subject

TABLE 4. Results for action recognition on UTD-MHAD dataset.

Method	Year	CS
STAR-Net [32]	2019	90.0
HDM-GB [33]	2019	92.8
PoseMap [29]	2018	94.5
HAMLET [28]	2020	95.1
BPAN [34]	2021	95.1
Ours	-	**96.3**

(CS). STAR-Net [32] uses 2D key-points from heatmap as input data. HDM-GB [33] uses RGB videos as input data. PoseMap [29] uses 3D pose from Kinect and heatmap as multi-modal input data. HAMLET [28] and BPAN [34] use RGB videos and skeletons from Kinect as multi-modal input data. Our method uses our skeletal representation as input and outperforms the other state-of-the-art methods.

V. CONCLUSION

In this paper, we proposed to learn the proper skeletal representation for the human action recognition problem. Especially, we constituted the overall framework by first reconstructing the 3D mesh vertices from the RGB video, then learning to sample the proper skeletal representation to improve the action recognition framework. From the experimental analysis, we verified two things: 1) By using the learned skeletal representation for action recognition, we confirmed on average around 6% accuracy improvement over the same SGN action classifier based on the sensor-captured skeletons, proving that there is much room for improvement in the sensor-captured skeletons, 2) We also obtained ablative results by varying the number of skeletons. Depending on data and protocols, the best accuracy has been obtained for around 40 and 80 skeletons. This suggests that the skeletal representation needs to be more densely sampled to model the human actions properly.

APPENDIX

MORE VISUALIZATION FOR SAMPLED SKELETONS AND IMPORTANCE FOR ACTION CLASSES

In Fig. 7, we visualized more examples for comparing the sensor-captured skeletons obtained from the Kinect sensor and our sampled skeletal representation from the transformer: We visualized the sensor-captured skeletons and our skeletal representation obtained from the transformer for individual action frames. Samples in the left columns are from the SYSU dataset, while the samples in the right columns are from the UTD dataset. Green and red dots represent the top-5 important skeletons for revealing the action classes obtained by the spatial max-pooling layer of the SGN action classifier [23] for sensor-captured skeletons and our sampled skeletons, respectively. We also visualize green and red dots in the same image frame to show their difference in 3rd and 6th columns. We can observe that red dots tend to vary depending on the different action classes in the dataset, while green dots tend to remain fixed regardless of the action classes in each dataset.

REFERENCES

[1] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Trear: Transformer-based RGB-D egocentric action recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 1, pp. 246–252, Mar. 2022.

[2] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proc. CVPR*, Jun. 2020, pp. 122–132.

[3] X. Qin, Y. Ge, J. Feng, D. Yang, F. Chen, S. Huang, and L. Xu, "DTMMN: Deep transfer multi-metric network for RGB-D action recognition," *Neurocomputing*, vol. 406, pp. 127–134, Sep. 2020.

[4] H. Wang, Z. Song, W. Li, and P. Wang, "A hybrid network for large-scale action recognition from RGB and depth modalities," *Sensors*, vol. 20, no. 11, p. 3305, Jun. 2020.

[5] J. Li, X. Xie, Q. Pan, Y. Cao, Z. Zhao, and G. Shi, "SGM-Net: Skeleton-guided multimodal network for action recognition," *Pattern Recognit.*, vol. 104, Aug. 2020, Art. no. 107356.

[6] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. CVPR*, Jun. 2015, pp. 1110–1118.

[7] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI*, 2018.

[8] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI*, 2017.

[9] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *Proc. CVPR*, Jun. 2018, pp. 5137–5146.

[10] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, Jun. 2011, pp. 1297–1304.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[12] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 2640–2649.

[13] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. CVPR*, Jun. 2021, pp. 16105–16114.

[14] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. CVPR*, Jun. 2018, pp. 7122–7131.

[15] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. CVPR*, Jun. 2019, pp. 10975–10985.

[16] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *Proc. ECCV*, 2020, pp. 20–40.

[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Nov. 2015.

[18] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. CVPR*, Jun. 2015, pp. 5344–5352.

[19] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. ICIP*, Sep. 2015, pp. 168–172.

[20] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proc. ICCV*, Oct. 2019, pp. 2252–2261.

[21] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. CVPR*, Jun. 2020, pp. 5253–5263.

[22] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. CVPR*, Jun. 2021, pp. 1954–1963.

[23] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. CVPR*, Jun. 2020, pp. 1112–1121.

[24] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proc. ECCV*, 2020, pp. 72–90.

[25] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. ICCV*, Oct. 2017, pp. 2117–2126.

[26] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.
[27] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "EleAtt-RNN: Adding attentiveness to neurons in recurrent neural networks," *IEEE Trans. Image Process.*, vol. 29, pp. 1061–1073, 2020.
[28] M. M. Islam and T. Iqbal, "HAMLET: A hierarchical multimodal attention-based human activity recognition algorithm," in *Proc. IROS*, Oct. 2020, pp. 10285–10292.
[29] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proc. CVPR*, Jun. 2018, pp. 1159–1168.
[30] H. Wang, B. Yu, J. Li, L. Zhang, and D. Chen, "Multi-stream interaction networks for human action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3050–3060, May 2022.
[31] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Learning latent global network for skeleton-based action prediction," *IEEE Trans. Image Process.*, vol. 29, pp. 959–970, 2020.
[32] W. McNally, A. Wong, and J. McPhee, "STAR-Net: Action recognition using spatio-temporal activation reprojection," in *Proc. CRV*, May 2019, pp. 49–56.
[33] R. Zhao, W. Xu, H. Su, and Q. Ji, "Bayesian hierarchical dynamic model for human action recognition," in *Proc. CVPR*, Jun. 2019, pp. 7733–7742.
[34] X. Weiyao, W. Muqing, Z. Min, and X. Ting, "Fusion of skeleton and RGB features for RGB-D human action recognition," *IEEE Sensors J.*, vol. 21, no. 17, pp. 19157–19164, Sep. 2021.
[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020, pp. 213–229.
[36] O. Dovrat, I. Lang, and S. Avidan, "Learning to sample," in *Proc. CVPR*, 2019, pp. 2760–2769.
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[38] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proc. CVPR*, Jun. 2019, pp. 558–567.

JUNUK CHA received the B.S. degree from the School of Mechanical Engineering, Ulsan National Institute of Science and Technology, Republic of Korea, in 2021. He is currently pursuing the combined master's and Ph.D. degree with the Ulsan National Institute of Science and Technology. His research interests include deep learning, computer vision, human pose and shape estimation, and human action recognition.

MUHAMMAD SAQLAIN received the B.S. degree (Hons.) in software engineering from Government College University Faisalabad, Pakistan, in 2014, the master's (M.S.) degree in computer software engineering from the National University of Science and Technology, Pakistan, in 2016, and the Ph.D. degree in computer science from Chungbuk National University, Republic of Korea, in 2021. He is currently a Postdoctoral Research Associate with the Ulsan National Institute of Science and Technology, Republic of Korea. His research interests include deep learning, fault detection, computer vision, and human action recognition.

DONGUK KIM received the B.S. degree in computer engineering from Yeongnam University, Republic of Korea, in 2021. He is currently pursuing the master's degree with the Department of Artificial Intelligence, Ulsan National Institute of Science and Technology. His research interests include pose estimation and 3D reconstruction.

SEUNGEUN LEE received the B.S. degree in information and communication engineering from Inha University, Republic of Korea, in 2021. He is currently pursuing the master's degree with the Department of Computer Science Engineering, Ulsan National Institute of Science and Technology. His research interests include pose estimation and 3D reconstruction.

SEONGYEONG LEE received the B.S. degree in IT convergence and application engineering in computer engineering from Pukyong National University, Busan, Republic of Korea, in 2020. She is currently pursuing the master's degree with the Department of Computer Science and Engineering, Ulsan National Institute of Science and Technology. Her research interests include pose estimation and face recognition.

SEUNGRYUL BAEK received the B.S. and M.S. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea, in 2009 and 2011, respectively, and the Ph.D. degree in electrical and electronic engineering from Imperial College London, U.K., in 2020. He is currently an Assistant Professor jointly affiliated to the AI Graduate School and the Department of Computer Science and Engineering, Ulsan National Institute of Science and Technology (UNIST), Republic of Korea. His research interests include deep learning, machine learning, and computer vision applications.