Skeleton Cloud Colorization for Unsupervised 3D Action Representation Learning

Siyuan Yang[1] Jun Liu[2*] Shijian Lu[1] Meng Hwa Er[1] Alex C. Kot[1]
[1]Nanyang Technological University [2]Singapore University of Technology and Design
siyuan005@e.ntu.edu.sg jun_liu@sutd.edu.sg {Shijian.Lu, emher, eackot}@ntu.edu.sg

Abstract

Skeleton-based human action recognition has attracted increasing attention in recent years. However, most of the existing works focus on supervised learning which requiring a large number of annotated action sequences that are often expensive to collect. We investigate unsupervised representation learning for skeleton action recognition, and design a novel skeleton cloud colorization technique that is capable of learning skeleton representations from unlabeled skeleton sequence data. Specifically, we represent a skeleton action sequence as a 3D skeleton cloud and colorize each point in the cloud according to its temporal and spatial orders in the original (unannotated) skeleton sequence. Leveraging the colorized skeleton point cloud, we design an auto-encoder framework that can learn spatial-temporal features from the artificial color labels of skeleton joints effectively. We evaluate our skeleton cloud colorization approach with action classifiers trained under different configurations, including unsupervised, semi-supervised and fully-supervised settings. Extensive experiments on NTU RGB+D and NW-UCLA datasets show that the proposed method outperforms existing unsupervised and semi-supervised 3D action recognition methods by large margins, and it achieves competitive performance in supervised 3D action recognition as well.

1. Introduction

Human action recognition is a fast developing area due to its wide applications in human-computer interaction, video surveillance, game control, etc. According to the types of input data, human action recognition can be grouped into different categories such as RGB-based [2, 35, 38, 42, 47, 49, 55, 56], depth-based [26, 28, 45] and 3D skeleton-based [1, 9, 17, 18, 20, 21, 31], etc. Among these types of inputs, 3D skeleton data, which represents a human body by the locations of keypoints in the 3D space, has attracted increasing attention in recent years. Compared with RGB videos or depth data, 3D skeleton data encodes high-level representations of human behaviors and it is generally lightweight and robust to the variations in appearances, surrounding distractions, viewpoint changes, etc. Additionally, skeleton sequences can be easily captured by depth sensors and thus a large number of supervised methods have been designed to learn spatio-temporal representations for skeleton based action recognition.

Deep neural networks have been widely studied to model the spatio-temporal representation of skeleton sequences under supervised scenarios [9, 20, 21, 31]. For example, Recurrent Neural Network (RNN) has been explored for modelling skeleton actions since it can capture temporal relations well [6, 20, 21, 29, 52]. Convolutional Neural Networks (CNNs) have also been explored to build skeleton-based recognition frameworks by converting joint coordinates to 2D maps [5, 9, 13, 43]. In addition, Graph Convolution Networks (GCNs) have attracted increasing attention due to their outstanding performance [27, 31, 48]. However, all these methods are supervised which require a large number of labelled training samples that are often costly to collect. How to learn effective feature representations with minimal annotations thus becomes critically important. To the best of our knowledge, only a few works [10, 14, 19, 37, 54] explored representation learning from unlabelled skeleton data for the task of action recognition, where the major approach is to reconstruct skeleton sequences from the encoded features via certain encoder-decoder structures. Unsupervised skeleton-based action representation learning remains a great challenge.

In this work, we propose to represent skeleton sequences as a 3D skeleton cloud, and design an unsupervised representation learning scheme that learns features from spatial and temporal color labels. We treat a skeletal sequence as a spatial-temporal skeleton cloud by stacking the skeleton data of all frames together, and colorize each point in the cloud according to its temporal and spatial orders in the original skeleton sequence. Specifically, we learn spatial-temporal features from the corresponding joints' colors by

[*] Corresponding author.

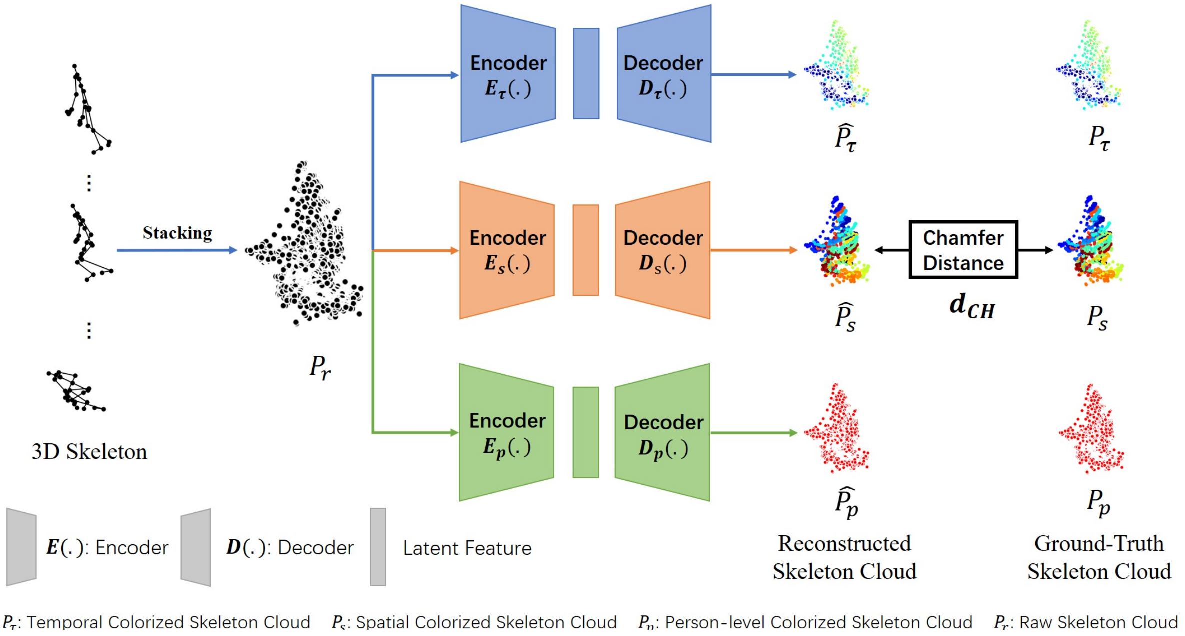

Figure 1. The pipeline of our proposed unsupervised representation learning method, using a novel skeleton cloud colorization scheme. Given a 3D skeleton sequence, we first stack it into a raw skeleton cloud P_r and then colorize it into 3 skeleton clouds P_τ, P_s, and P_p (construction details shown in Figs. 3 and 4) according to spatial, temporal, and person-level information, respectively. With the three colorized clouds as self-supervision signal, three encoder-decoders (with the same structure but no weight sharing) learn discriminative skeleton representative features. (The encoder and decode details are provided in supplementary material.)

leveraging a point-cloud based auto-encoder framework as shown in Fig. 1. By repainting the whole skeleton cloud, our network can achieve unsupervised skeleton representation learning successfully by learning both spatial and temporal information from unlabeled skeleton sequences.

The contributions of this paper are threefold. *First*, we formulate unsupervised action representation learning as 3D skeleton cloud repainting problem, where each skeleton sequence is treated as a skeleton cloud and can be directly processed with a point cloud auto-encoder framework. *Second*, we propose a novel skeleton cloud colorization scheme that colorizes each point in the skeleton cloud based on its temporal and spatial orders in the skeleton sequence. The color labels 'fabricate' self-supervision signals which boost unsupervised skeleton action representation learning significantly. *Third*, extensive experiments show that our method outperforms state-of-the-art unsupervised and semi-supervised skeleton action recognition methods by large margins, and its performance is also on par with supervised skeleton-based action recognition methods.

To the best of our knowledge, this is the first work that converts the unsupervised skeleton representation learning problem into a novel skeleton cloud repainting task.

2. Related work

Skeleton-based action recognition. Skeleton-based action recognition has attracted increasing interest recently. Unlike traditional methods that design hand-craft features [8, 39, 40, 46], deep-learning based methods employ Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Graph Convolution Networks (GCNs) to learn skeleton-sequence representation directly. Specifically, RNNs have been widely used to model temporal dependencies for skeleton-based action recognition. For example, [6] uses a hierarchical RNN model to represent human body structures and temporal dynamics of the body joints. [20, 21] proposes a 2D Spatio-Temporal LSTM framework to employ the hidden sources of action related information over both spatial and temporal domains concurrently. [52] adds a view-adaptation scheme to the LSTM to regulate the observation viewpoints.

CNN-based methods [5, 9, 13, 43] have also been proposed for skeleton action recognition. They usually transform the skeleton sequences to skeleton maps of same target size and then use CNNs to learn the spatial and temporal dynamics. For example, [5, 13] transform a skeleton sequence to an image by treating the joint coordinate (x,y,z) as the R, G, and B channels of a pixel. [9] transforms the 3D skeleton data to three skeleton clips for robust action feature learning. [43] presents a "scene flow to action map" representation for action recognition with CNNs.

Inspired by the observation that human 3D-skeleton is naturally a topological graph, Graph Convolutional

Networks (GCN) have attracted increasing attention in skeleton-based action recognition. For example, [48] presents a spatial-temporal GCN to learn both spatial and temporal patterns from skeleton data. [31] uses a non-local method with the spatial-temporal GCN to boost performance. [30] uses bone features for skeleton-based action recognition. [27] recognizes actions by searching for different graphs at different layers via neural architecture search.

Though the aforementioned methods achieve very impressive performance, they are all supervised requiring large amount of labelled data which is prohibitively time-consuming to collect. In this work, we study unsupervised representation learning in skeleton-based action recognition which mitigates the data labelling constraint greatly.

Unsupervised representation learning for action recognition. Unsupervised action recognition aims to learn effective feature representations by predicting future frames of input sequences or by re-generating the sequences. Most existing methods focus on RGB videos or RGB-D videos. For example, [36] uses a LSTM based Encoder-Decoder architecture to learn video representations [23] uses a RNN based encoder-decoder framework to predict the sequences of flows computed with RGB-D modalities. [15] uses unlabeled video to learn view-invariant video representations.

Unsupervised skeleton-based action recognition was largely neglected though a few works attempt to address this challenging task very recently. For example, [54] presents a GAN encoder-decoder to re-generate masked input sequences. [10] adopts a hierarchical fusion approach to improve human motion generation. [37] presents a decoder-weakening strategy to drive the encoder to learn discriminative action features.

The aforementioned methods all process skeleton sequences frame by frame and extract temporal features from ordered sequences. We instead treat a skeleton sequence as a novel colored skeleton cloud by stacking human joints of each frame together. We design a novel skeleton colorization scheme and leverage the color information for unsupervised spatial-temporal representation learning.

3. Method

In this section, we present our skeleton cloud colorization representation learning method that converts the skeleton sequence to a skeleton cloud, and colorizes each point in the cloud by its spatial-temporal properties. In particular, we present how to construct the skeleton cloud in Section 3.1 and describe the colorization step in Section 3.2. Repainting pipeline and training details are discribed in Section 3.3 and Section 3.4, respectively.

3.1. Data Processing

Given a skeleton sequence S under the global coordinate system, the j^{th} skeleton joint in the t^{th} frame is denoted as $v_{t,j} = [x_{t,j}, y_{t,j}, z_{t,j}]$, $t \in (1, \cdots, T)$, $j \in (1, \cdots, J)$, where T and J denote the number of frames and body joints, respectively. Generally, skeleton data is defined as a sequence and the set of joints in the t^{th} frame are denoted as $V_t = \{v_{t,j} | j = 1, ..., J\}$. We propose to treat all the joints in a skeleton sequence as a whole by stacking all frames' data together, and Fig. 1 illustrates the stacking framework. We name the stacked data as skeleton cloud and denote it by $P_r = \{v_{t,j} = [x_{t,j}, y_{t,j}, z_{t,j}] | t = 1, ..., T; j = 1, ..., J\}$. The obtained 3D skeleton cloud therefore consists of $N = T \times J$ 3D points in total. We use P_r to denote raw skeleton cloud so as to differentiate it from the colorized clouds to be described later.

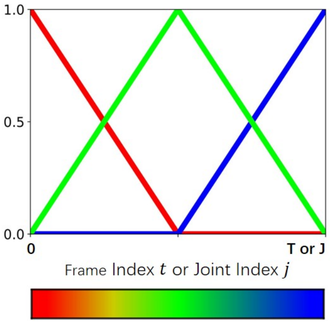

Figure 2. Illustration of the colorization scheme for temporal index t and spatial index j. Top: Definition of each color channel(RGB) when varying t or j (where $t \in [1, T]; j \in [1, J]$). Bottom: Corresponding color of the temporal index t and spatial index j. With the increase of the point's temporal/spatial order index, the corresponding color changes from red to green to blue. (Best viewed in color)

3.2. Skeleton Cloud Colorization

Points within our skeleton cloud are positioned with 3D spatial coordinates (x, y, z) which is similar to normal point cloud that consists of unordered points. The spatial relation and temporal dependency of skeleton cloud points are crucial in skeleton-based action recognition, but they are largely neglected in the aforementioned raw skeleton cloud data. We propose an innovative skeleton cloud colorization method to exploit the spatial relation and temporal dependency of skeleton cloud points for unsupervised skeleton-based action representation learning.

Temporal Colorization: Temporal information is critical in action recognition. To assign each point in the skeleton cloud a temporal feature, we colorize the skeleton cloud points according to their relative time order (from 1 to T) in the original skeleton sequence. Different colorization schemes have been reported and here we adopt the colorization scheme that uses 3 RGB channels [4] as illustrated in Fig. 2. The generated color is actually relatively linear under this color scheme. Hence, points from adjacent frames are assigned with similar color under this distribu-

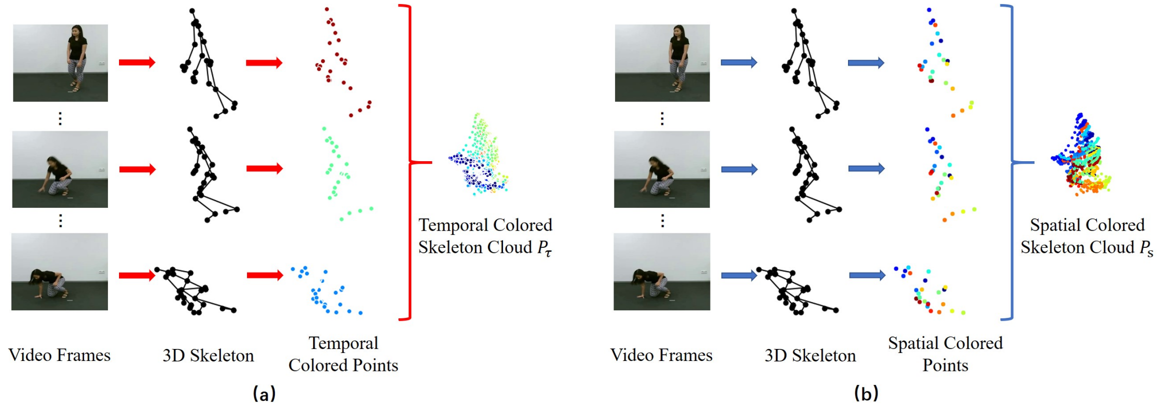

Figure 3. The pipelines of temporal colorization and spatial colorization. (a) Given a skeleton sequence, the temporal colorization colorizes points based on the relative temporal order t ($t \in [1, T]$) in the sequential data. (b) The spatial colorization colorizes points based on the index of joints j ($j \in [1, J]$). (Best viewed in color)

tion, which facilitates the learning of temporal order information. The value distributions of R, G, B channels can be formulated as follows:

$$r^{\tau}_{t,j} = \begin{cases} -2 \times (t/T) + 1, & \text{if } t <= T/2 \\ 0, & \text{if } t > T/2 \end{cases} \tag{1}$$

$$g^{\tau}_{t,j} = \begin{cases} 2 \times (t/T), & \text{if } t <= T/2 \\ -2 \times (t/T) + 2, & \text{if } t > T/2 \end{cases} \tag{2}$$

$$b^{\tau}_{t,j} = \begin{cases} 0, & \text{if } t <= T/2 \\ 2 \times (t/T) - 1, & \text{if } t > T/2 \end{cases} \tag{3}$$

With this colorizing scheme, we can assign different colors to points from different frames based on the frame index t as illustrated in Fig. 3(a). More specifically, with this temporal-index based colorization scheme, each point will have a 3-channels features that can be visualized with red, green and blue channels (RGB channels) to represent its temporal information. Together with the original 3D coordinate information, the temporally colorized skeleton cloud can be denoted by $P_\tau = \{v^{\tau}_{t,j} = [x_{t,j}, y_{t,j}, z_{t,j}, r^{\tau}_{t,j}, g^{\tau}_{t,j}, b^{\tau}_{t,j}] | t = 1, ..., T; j = 1, ..., J\}$.

Spatial Colorization: Besides temporal information, spatial information is also very important for action recognition. We employ similar colorization scheme to colorize spatial information as illustrated in Fig. 2. The scheme assigns different colors to different points according to their spatial orders $j \in [1, J]$ (J is the total number of joints in the skeleton cloud of a person), as shown in Figure 3(b). The distribution of the values of R, G, B channels can be calculated as follows:

$$r^{s}_{t,j} = \begin{cases} -2 \times (j/J) + 1, & \text{if } j <= J/2 \\ 0, & \text{if } j > J/2 \end{cases} \tag{4}$$

$$g^{s}_{t,j} = \begin{cases} 2 \times (j/J), & \text{if } j <= J/2 \\ -2 \times (j/J) + 2, & \text{if } j > J/2 \end{cases} \tag{5}$$

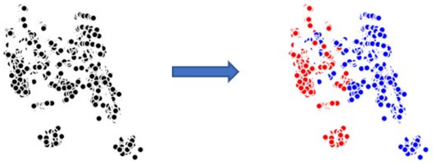

Figure 4. Person-level colorization. The first person's points will be assigned the red color, and person two will colorize to blue.

$$b^{s}_{t,j} = \begin{cases} 0, & \text{if } j <= J/2 \\ 2 \times (j/J) - 1, & \text{if } j > J/2 \end{cases} \tag{6}$$

We denote the spatially colorized skeleton cloud as $P_s = \{v^{s}_{t,j} = [x_{t,j}, y_{t,j}, z_{t,j}, r^{s}_{t,j}, g^{s}_{t,j}, b^{s}_{t,j}] | t = 1, ..., T; j = 1, ..., J\}$. With the increase of the spatial order index of the joint in the skeleton, points will be assigned with different colors that change from red to blue and to green gradually.

Person-level Colorization: Human actions contain rich person interaction information as in NTU RGB+D [29] which is important to the skeleton action recognition. We therefore propose a person-level colorization scheme for action recognition.

We focus on the scenarios that human interactions involve two persons only, and apply different colors to the points of different persons. Specifically, we encode the first person's joints with red color and the second person's joints with blue color as illustrated in Fig. 4. The person-level colored clouds can thus be denoted by $P_p = \{v^{p}_{t,j,n} = [x_{t,j,n}, y_{t,j,n}, z_{t,j,n}, 1, 0, 0] | t = 1, ..., T; j = 1, ..., J; n = 1\} \cup \{v^{p}_{t,j,n} = [x_{t,j,n}, y_{t,j,n}, z_{t,j,n}, 0, 0, 1] | t = 1, ..., T; j = 1, ..., J; n = 2\}$,where $n = 1$ and $n = 2$ mean that the points belong to the first and the second persons, respectively.

Given a raw skeleton cloud, the three colorization schemes thus construct three colorized skeleton clouds P_τ, P_s and P_p that capture temporal dependency, spatial relations and human interaction information, respectively.

3.3. Repainting Pipeline

Inspired by the success of self-supervised learning, our goal is to extract the temporal, spatial, and interactive information by learning to repaint the raw skeleton cloud P_r in self-supervised manner. As illustrated in Fig. 1, we use colorized skeleton clouds (temporal-level P_τ, spatial-level P_s, and person-level P_p) as three kinds of self-supervision signals, respectively. The framework consists of an encoder $E(.)$ and a decoder $D(.)$. Since we have three colorization schemes, we have three pairs of encoders ($E_\tau(.)$, $E_s(.)$, and $E_p(.)$) and decoders ($D_\tau(.)$, $D_s(.)$, and $D_p(.)$). Below we use the temporal colorization stream as the example to explain the model architecture and the training process.

Model Architecture: As mentioned in Section 3.2, the obtained skeleton cloud format is similar to that of the normal point cloud. We therefore adopt DGCNN [44] (designed for point cloud classification and segmentation) as the backbone of our framework and use the modules before the fully-connected (FC) layers to build our encoder [1].

In addition, we adopt the decoder [1] of FoldingNet [50] as the decoder of our network architecture. Since the input and output of FoldingNet are all $N \times 3$ matrices with 3D positions (x, y, z) only, we enlarge the feature dimension to 6 to repaint both the position and color information. Assuming that the input is the raw point set P_r and the obtained repainted point set is $\widehat{P_\tau} = D_\tau(E_\tau(P_r))$, the repainting error between the ground truth temporal colorization P_τ and the repainted $\widehat{P_\tau}$ is computed by using the Chamfer distance:

$$d_{CH}(P_\tau, \widehat{P_\tau}) = Max\left\{A, B\right\}, where \tag{7}$$

$$A = \frac{1}{|P_\tau|} \sum_{v_\tau \in P_\tau} \min_{\widehat{v_\tau} \in \widehat{P_\tau}} \|v_\tau - \widehat{v_\tau}\|_2 \tag{8}$$

$$B = \frac{1}{\left|\widehat{P_\tau}\right|} \sum_{\widehat{v_\tau} \in \widehat{P_\tau}} \min_{v_\tau \in P_\tau} \|\widehat{v_\tau} - v_\tau\|_2 \tag{9}$$

where the term $\min\limits_{\widehat{v_\tau} \in \widehat{P_\tau}} \|v_\tau - \widehat{v_\tau}\|_2$ enforces that any 3D point v_τ in temporally colorized skeleton cloud P_τ has a matched 3D point $\widehat{v_\tau}$ in the repainted point cloud $\widehat{P_\tau}$. The term $\min\limits_{v_\tau \in P_\tau} \|\widehat{v_\tau} - v_\tau\|_2$ enforces the matching vice versa. The max operation enforces that the distance from P_τ to $\widehat{P_\tau}$ and vice versa need to be small concurrently.

By using the Chamfer distance, the encoder $E_\tau(.)$ and decoder $D_\tau(.)$ are forced to learn temporal dependency via the proposed temporal repainting scheme that reconstructs temporal-order colors. Similarly, the encoder $E_s(.)$ and decoder $D_s(.)$ will learn useful spatial relations during spatial repainting, and the encoder $E_p(.)$ and decoder $D_p(.)$ are pushed to distinguish the person index and learn interactive information. It is non-trivial to repaint P_r to colorized skeleton clouds. To balance the color repainting and unsupervised feature learning, we uniformly sample half of points in P_r for colorization. Specifically, in the temporal colorization stream, points corresponding to odd-order frames are colored, while the rest is not. In the spatial colorization, points with odd-index joints are colored based on joints orders, and even-indexed joints are not colored.

Table 1. Comparisons to state-of-the-art unsupervised skeleton action recognition method on NTU RGB+D dataset. The evaluation setting is as in [10, 19, 25, 37, 54]. ('*TS*':Temporal Stream; '*SS*':Spatial Stream; '*PS*':Person Stream)

Method	**NTU RGB+D**	
	C-Subject	C-View
LongT GAN [54]	39.1	52.1
MS^2L [19]	52.6	–
P&C FS-AEC [37]	50.6	76.3
P&C FW-AEC [37]	50.7	76.1
EnGAN-PoseRNN [10]	68.6	77.8
SeBiReNet [25]	–	79.7
'*TS*' Colorization (Ours)	71.6	79.9
'*TS* + *SS*' Colorization (Ours)	74.6	82.6
'*TS* + *SS* + *PS*' Colorization (Ours)	**75.2**	**83.1**

Table 2. Comparisons to state-of-the-art unsupervised skeleton action recognition method on NW-UCLA dataset. The evaluation setting is as in [10, 19, 25, 37, 54].

Method	**NW-UCLA**
LongT GAN [54]	74.3
MS^2L [19]	76.8
SeBiReNet [25]	80.3
P&C FS-AEC [37]	83.8
P&C FW-AEC [37]	84.9
'*TS*' Colorization (Ours)	90.1
'*TS* + *SS*' Colorization (Ours)	**91.1**

3.4. Training for Skeleton Action Recognition

After the self-supervised repainting, we obtain three encoders (*i.e.*, $E_\tau(.)$. $E_s(.)$, and $E_p(.)$) that capture meaningful temporal, spatial, interaction features, respectively. With the feature representations from the three encoders, we include a simple linear classifier $f(.)$ on top of the encoder to perform action recognition as in [10, 19, 25, 37, 54]. We adopt different settings to train the classifier including unsupervised, semi-supervised, and supervised settings. In the unsupervised setting, the encoder is only trained by the skeleton cloud repainting method, and then we train the linear classifier with the encoder fixed by following previous unsupervised skeleton representation learning works [10, 19, 25, 37, 54]. In the semi-supervised and supervised settings, the encoder is first trained with unsupervised rep-

[1] Detailed structure of encoder and decoder can be found in supplementary.

Table 3. Comparisons of action recognition results with semi-supervised learning approaches on NTU RGB+D dataset. The number in parentheses denotes the number of labeled samples per class.

Method	1%		5%		10%		20%		40%	
	CS (7)	CV (7)	CS (33)	CV (31)	CS (66)	CV (62)	CS (132)	CV (124)	CS (264)	CV (248)
Pseudolabels [11]	–	–	50.9	56.3	57.2	63.1	62.4	70.4	68.0	76.8
VAT [24]	–	–	51.3	57.9	60.3	66.3	65.6	72.6	70.4	78.6
VAT + EntMin	–	–	51.7	58.3	61.4	67.5	65.9	73.3	70.8	78.9
S^4L(Inpainting) [51]	–	–	48.4	55.1	58.1	63.6	63.1	71.1	68.2	76.9
ASSL [34]	–	–	57.3	63.6	64.3	69.8	68.0	74.7	72.3	80.0
LongT GAN [54]	35.2	–	–	–	62.0	–	–	–	–	–
MS^2L [19]	33.1	–	–	–	65.2	–	–	–	–	–
'*TS*' Colorization (Ours)	42.9	46.3	60.1	63.9	66.1	73.3	72.0	77.9	75.9	82.7
'*TS* + *SS*' Colorization (Ours)	48.1	51.5	64.7	69.3	70.8	78.2	75.2	81.8	79.2	86.0
'*TS* + *SS* + *PS*' Colorization (Ours)	**48.3**	**52.5**	**65.7**	**70.3**	**71.7**	**78.9**	**76.4**	**82.7**	**79.8**	**86.8**

Table 4. Comparisons of action recognition results with semi-supervised learning approaches on NW-UCLA dataset. $v./c.$ denotes the number of labeled videos per class.

Method	1% (1 $v./c.$)	5% (5 $v./c.$)	10% (10 $v./c.$)	15% (15 $v./c.$)	30% (30 $v./c.$)	40% (40 $v./c.$)
Pseudolabels [11]	–	35.6	–	48.9	60,6	65.7
VAT [24]	–	44.8	–	63.8	73.7	73.9
VAT + EntMin [7]	–	46.8	–	66.2	75.4	75.6
S^4L(Inpainting) [51]	–	35.3	–	46.6	54.5	60.6
ASSL [34]	–	52.6	–	74.8	78.0	78.4
LongT GAN [54]	18.3	–	59.9	–	–	–
MS^2L [19]	21.3	–	60.5	–	–	–
'*TS*' Colorization (Ours)	40.6	55.9	71.3	74.3	81.4	83.6
'*TS* + *SS*' Colorization (Ours)	**41.9**	**57.2**	**75.0**	**76.0**	**83.0**	**84.9**

resentation learning and then fine-tuned with the linear classifier as in [19]. We use the standard cross-entropy loss as the classification loss L_{cls}.

4. Experiments

We conducted extensive experiments over two publicly accessible datasets including NTU RGB+D [29], and Northwestern-UCLA [41]. The experiments aim to evaluate whether our skeleton cloud colorization scheme can learn effective unsupervised feature representations for the task of skeleton action recognition. We therefore evaluate different experimental settings including unsupervised and semi-supervised as well as supervised.

4.1. Datasets

NTU RGB+D [29]: NTU RGB+D consists of 56880 skeleton action sequences which is the most widely used dataset for skeleton-based action recognition research. In this dataset, action samples are performed by 40 volunteers and categorized into 60 classes. Each sample contains an action and is guaranteed to have at most two subjects as captured by three Microsoft Kinect v2 cameras from different views. The authors of this dataset recommend two benchmarks: (1) cross-subject (CS) benchmark where training data comes from 20 subjects and testing data comes from the other 20 objects; (2) cross-view (CV) benchmark where training data comes from camera views 2 and 3, and testing data comes from camera view 1.

Northwestern-UCLA (NW-UCLA) [41]: This dataset is captured by Kinect v1 cameras and it contains 1494 samples performed by 10 volunteers. It contains 10 action classes, and each body has 20 skeleton joints. Following the evaluation protocol in [41], the training set consists of samples from camera views 1 and 2, and the rest samples from camera view 3 form the tesing set.

4.2. Implementation Details

For NTU RGB+D, we adopt the pre-processing in [31] and uniformly sampled $T = 40$ frames from each skeleton sequence. The sampled skeleton sequences are constructed to a 2000-points skeleton cloud. For NW-UCLA, we adopt the pre-processing in [37] and uniformly sampled $T = 50$ frames from the skeleton sequences. The skeleton cloud has 1000 points.

In the unsupervised feature learning phase, we use Adam optimizer and set the initial learning rate at 1e-5 and reduce it to 1e-7 with cosine annealing. The dimension of the encoder output is 1024, and the batch size is 24 for both datasets. The training lasts for 150 epochs. In the classifier training phase, we use SGD optimizer with Nesterov momentum (0.9). We set the initial learning rate at 0.001 and reduce it to 1e-5 with cosine annealing. The batch size for NTU RGB+D and NW-UCLA is 32 and 64 and the training lasts for 100 epochs. We implement our method using PyTorch and all experiments were conducted on Tesla P100 GPUs with CUDA 10.1.

Table 5. Comparisons to state-of-the-art supervised skeleton action recognition method on NTU RGB+D dataset.

Method	NTU RGB+D	
	C-Subject	C-View
Supervised method		
ST-LSTM [21]	69.2	77.7
GCA-LSTM [22]	74.4	82.8
ST-GCN [48]	81.5	88.3
AS-GCN [16]	86.8	94.2
2s AGC-LSTM [33]	89.2	95.0
4s MS-AAGCN [32]	90.0	96.2
4s Shift-GCN [3]	**90.7**	**96.5**
Unsupervised Pretrain		
Li *et al.* [15]	63.9	68.1
MS^{2}L [19]	78.6	–
‘TS’ Colorization (Ours)	84.2	93.1
‘$TS + SS$’ Colorization (Ours)	86.3	94.2
‘$TS + SS + PS$’ Colorization (Ours)	**88.0**	**94.9**

Table 6. Comparisons to state-of-the-art supervised skeleton action recognition method on NW-UCLA dataset.

Method	NW-UCLA
Supervised method	
Actionlet ensemble [40]	76.0
HBRNN-L [6]	78.5
Ensemble TS-LSTM [12]	89.2
VA-RNN-Aug [53]	90.7
2s AGC-LSTM [33]	93.3
1s Shift-GCN [3]	92.5
4s Shift-GCN [3]	**94.6**
Unsupervised Pretrain	
Li *et al.* [15]	62.5
MS^{2}L [19]	86.8
‘TS’ Colorization (Ours)	92.7
‘$TS + SS$’ Colorization (Ours)	**94.0**

Table 7. Comparisons of different network configurations' results with unsupervised and supervised setting on NTU RGB+D and NW-UCLA datasets.

Dataset	NTU-CS	NTU-CV	NW-UCLA
Unsupervised Setting			
Baseline-U	61.8	68.4	78.6
‘TS Colorization’	71.6	79.9	90.1
‘SS Colorization’	68.4	77.5	87.0
‘PS Colorization’	64.2	72.8	–
‘$TS + SS$’ Colorization	74.6	82.6	**91.1**
‘$TS + PS$’ Colorization	73.3	81.4	–
‘$JS + PS$’ Colorization	69.6	78.6	–
‘$TS + SS + PS$’ Colorization	**75.2**	**83.1**	–
Supervised Setting			
Baseline-S	76.5	83.4	83.8
‘TS’ Colorization	84.2	93.1	92.7
‘SS’ Colorization	82.3	91.5	90.4
‘PS’ Colorization	81.1	90.3	–
‘$TS + SS$’ Colorization	86.3	94.2	**94.0**
‘$TS + PS$’ Colorization	86.4	94.1	–
‘$SS + PS$’ Colorization	85.0	93.0	–
‘$TS + SS + PS$’ Colorization	**88.0**	**94.9**	–

4.3. Comparison with the state-of-the-art methods

We conduct extensive experiments under three settings including unsupervised learning, semi-supervised learning and supervised learning. We also study three skeleton cloud colorization configurations: 1) ‘*T-Stream (TS)*’ that uses temporally colorized skeleton cloud as self-supervision; 2) ‘*S-Stream (SS)*’ that uses spatially colorized skeleton cloud as self-supervision; and 3) ‘*P-Stream (PS)*’ that uses person-level colorized cloud as self-supervision.

Unsupervised Learning. In the unsupervised setting, the feature extractor (i.e. the encoder $E(.)$) is trained with our proposed skeleton cloud colorization unsupervised representation learning approach. Then the feature representation is evaluated by the simple linear classifier $f(.)$, which is trained on the top of the frozen encoders $E(.)$. Such experimental setting for unsupervised learning has been widely adopted and practised in prior studies [10, 19, 25, 37, 54]. Here for fair comparisons, we use the same setting as these prior works.

We compare our skeleton cloud colorization method with prior unsupervised methods on NTU RGB+D and NW-UCLA datasets as shown in Tables 1 and 2. It can be seen that our proposed temporal colorization encoding (i.e. *’TS’* colorization) outperforms prior unsupervised methods on NTU RGB+D dataset, especially under the cross-subject evaluation protocol. Additionally, the proposed ‘TS + SS’ colorization and ‘$TS + SS + PS$’ colorization outperform the state-of-the-art obviously on both cross subject and cross view protocols. For NW-UCLA, our method outperforms the state-of-the-art consistently for both configurations ‘TS’ and ‘$TS + SS$’ as shown in Table 2.

Semi-supervised Learning. We evaluate semi-supervised learning with the same protocol as in [19, 34] for fair comparisons. Under the semi-supervised setting, the encoder $E(.)$ is first pre-trained with colorized skeleton clouds and then jointly trained with the linear classifier $f(.)$ with a small ratio of action annotations. Following [19, 34], we derive labelled data by uniformly sampling 1%, 5%, 10%, 20%, 40% data from the training set of NTU RGB+D dataset, and 1%, 5%, 10%, 15%, 30%, 40% data from the training set of NW-UCLA dataset, respectively.

Tables 3 and 4 show experimental results on NTU RGB+D and NW-UCLA datasets, respectively. As Table 3 shows, our method performs better than the state-of-the-art consistently for all three configurations (‘TS’, ‘$TS + SS$’, and ‘$TS + SS + PS$’) on NTU RGB+D. For NW-UCLA dataset, our proposed temporal colorization encoding (‘TS’) performs better than the state-of-the-art as shown in Table 4. Additionally, our ‘$TS + SS$’ colorization scheme outperforms the state-of-the-art by large margins.

Supervised Learning. Following the supervised evalu-

Table 8. Comparisons of different network configurations' results with semi-supervised setting on NW-UCLA dataset.

Method	1% (1 $v./c.$)	5% (5 $v./c.$)	10% (10 $v./c.$)	15% (15 $v./c.$)	30% (30 $v./c.$)	40% (40 $v./c.$)
Baesline-Semi	34.3	46.4	54.9	61.8	69.1	70.2
'*TS*' Colorization	40.6	55.9	71.3	74.3	81.4	83.6
'*SS*' Colorization	39.1	54.2	66.3	70.2	79.1	80.8
'*TS* + *SS*' Colorization	**41.9**	**57.2**	**75.0**	**76.0**	**83.0**	**84.9**

Table 9. Comparisons of different network configurations' results with semi-supervised setting on NTU RGB+D dataset.

Method	1%		5%		10%		20%		40%	
	CS (7)	CV (7)	CS (33)	CV (31)	CS (66)	CV (62)	CS (132)	CV (124)	CS (264)	CV (248)
Baseline-Semi	27.1	28.1	46.0	50.6	55.1	60.7	60.9	69.1	64.2	73.7
'*TS*' Colorization	42.9	46.3	60.1	63.9	66.1	73.3	72.0	77.9	75.9	82.7
'*SS*' Colorization	40.2	43.1	54.6	60.0	60.1	68.1	64.2	73.1	69.1	77.6
'*PS*' Colorization	37.9	40.1	51.2	56.0	56.8	63.2	61.9	70.2	65.8	74.59
'*TS* + *SS*' Colorization	48.1	51.5	64.7	69.3	70.8	78.2	75.2	81.8	79.2	86.0
'*TS* + *PS*' Colorization	46.9	50.5	63.9	68.1	69.8	77.0	74.9	81.3	78.3	85.4
'*SS* + *PS*' Colorization	43.2	46.7	58.3	64.4	64.2	72.0	69.0	77.1	73.2	81.5
'*TS* + *SS* + *PS*' Colorization	**48.3**	**52.5**	**65.7**	**70.3**	**71.7**	**78.9**	**76.4**	**82.7**	**79.8**	**86.8**

ation protocol in [19], we pre-train the encoder with our unsupervised skeleton colorization method and fine-tune the encoder and classifier by using labeled training data. Tables 5 and 6 show experimental results. We can observe that our method achieves superior performance on NW-UCLA, and it performs much better than the previous "Unsupervised Pre-trained" methods [15, 19] (first train under the unsupervised feature learning and then fine-tune the framework with the labeled data). On the large-scale NTU RGB+D, our method outperforms the "Unsupervised Pre-trained" works [15, 19] by large margins. Though our framework is not designed for supervised setting, its performance is even comparable to state-of-the-art supervised methods.

4.4. Ablation Study

Effectiveness of Our Skeleton Colorization: We verify the effectiveness of our skeleton cloud colorization on all three learning settings including unsupervised learning, semi-supervised learning, and fully-supervised learning. We compare our method with three baselines: 1) *Baseline-U*: it only trains the linear classifier and freezes the encoder which is randomly initialized; 2) *Baseline-Semi*: the encoder is initialized with random weight instead of pre-training by our unsupervised representation learning; 3) *Baseline-S*: the same as *Baseline-Semi*. We train the encoder and linear classifier jointly with action labels. The input for these three baselines is the raw skeleton cloud without color label information.

Tables 7, 8 and 9 show experimental results. It can be seen that all the three colorization strategies (i.e. temporal-level, spatial-level, and person-level) achieve significant performance improvement as compared with the baseline, demonstrating the effectiveness of our proposed colorization technique. Though the person-level colorization stream does not perform as well as the other two streams on the NTU RGB+D, it improves the overall performance while collaboration with the other two.

Effectiveness of Colorization Ratio: As mentioned in Section 3.3, we need to balance between the repainting and the spatial-temporal ordered feature learning. We observe that the unsupervised performance improves from 65.7% to 71.6% on the temporal stream on the NTU RGB+D (Cross-subject setting) when 50% points with the color information is provided, demonstrating the effectiveness of our proposed colorization scheme. Detailed comparisons can be found in the supplementary materials.

5. Conclusion

In this paper, we address unsupervised representation learning in skeleton action recognition, and design a novel skeleton cloud colorization method that is capable of learning skeleton representations from unlabeled data. We obtain colored skeleton cloud representations by stacking skeleton sequences to 3D skeleton cloud and colorizing each point according to its temporal and spatial orders in the skeleton sequences. Additionally, spatio-temporal features are learned effectively from the corresponding joints' colors from unlabeled data. The experiments demonstrate that our proposed method achieves superior unsupervised action recognition performance.

Acknowledgement. This work was done at Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University. This research is supported in part by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation), the Science and Technology Foundation of Guangzhou Huangpu Development District under Grant 2019GH16, and China-Singapore International Joint Research Institute under Grant 206-A018001. This research work is also partially supported by the SUTD project PIE-SGP-Al-2020-02.

References

[1] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, and et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[3] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020. 7

[4] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018. 3

[5] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 579–583. IEEE, 2015. 1, 2

[6] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015. 1, 2, 7

[7] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 6

[8] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-third international joint conference on artificial intelligence*, 2013. 2

[9] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017. 1, 2

[10] Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and Venkatesh Babu Radhakrishnan. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467. IEEE, 2019. 1, 3, 5, 7

[11] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013. 6

[12] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1012–1020, 2017. 7

[13] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 597–600. IEEE, 2017. 1, 2

[14] Jingyuan Li and Eli Shlizerman. Sparse semi-supervised action recognition with active learning. *arXiv preprint arXiv:2012.01740*, 2020. 1

[15] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Unsupervised learning of view-invariant action representations. In *Advances in Neural Information Processing Systems*, pages 1254–1264, 2018. 3, 7, 8

[16] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3595–3603, 2019. 7

[17] Tianjiao Li, Qiuhong Ke, Hossein Rahmani, Rui En Ho, Henghui Ding, and Jun Liu. Else-net: Elastic semantic network for continual action recognition from skeleton data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 1

[18] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16266–16275, June 2021. 1

[19] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2490–2498, 2020. 1, 5, 6, 7, 8

[20] Jun Liu, Amir Shahroudy, Dong Xu, Alex C Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3007–3021, 2017. 1, 2

[21] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. 1, 2, 7

[22] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1647–1656, 2017. 7

[23] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2212, 2017. 3

[24] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 6

[25] Qiang Nie and Yunhui Liu. View transfer on human skeleton pose: Automatically disentangle the view-variant and view-invariant information for pose representation learning. *International Journal of Computer Vision*, pages 1–22, 2020. 5, 7

[26] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 716–723, 2013. 1

[27] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. Learning graph convolutional network for skeleton-based human action recognition by neural searching. *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, 2020. 1, 3

[28] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European conference on computer vision*, pages 742–757. Springer, 2014. 1

[29] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 4, 6

[30] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[31] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 3, 6

[32] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020. 7

[33] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1227–1236, 2019. 7

[34] Chenyang Si, Xuecheng Nie, Wei Wang, Liang Wang, Tieniu Tan, and Jiashi Feng. Adversarial self-supervised learning for semi-supervised 3d action recognition. In *European Conference on Computer Vision (ECCV)*, 2020. 6, 7

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 1

[36] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015. 3

[37] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020. 1, 3, 5, 6, 7

[38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[39] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014. 2

[40] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3d human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):914–927, 2013. 2, 7

[41] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014. 6

[42] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 1

[43] Pichao Wang, Wanqing Li, Zhimin Gao, Yuyao Zhang, Chang Tang, and Philip Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2017. 1, 2

[44] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 5

[45] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, and Junsong Yuan. 3dv: 3d dynamic voxel for action recognition in depth video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 511–520, 2020. 1

[46] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27. IEEE, 2012. 2

[47] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 1

[48] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI conference on artificial intelligence*, 2018. 1, 3, 7

[49] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis. In

European Conference on Computer Vision, pages 769–786. Springer, 2020. 1

[50] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018. 5

[51] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019. 6

[52] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2117–2126, 2017. 1, 2

[53] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1963–1978, 2019. 7

[54] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Thirty-Second AAAI conference on artificial intelligence*, 2018. 1, 3, 5, 6, 7

[55] Hongyuan Zhu, Romain Vial, and Shijian Lu. Tornado: A spatio-temporal convolutional regression network for video action proposal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5813–5821, 2017. 1

[56] Hongyuan Zhu, Romain Vial, Shijian Lu, Xi Peng, Huazhu Fu, Yonghong Tian, and Xianbin Cao. Yotube: Searching action proposal via recurrent and static regression networks. *IEEE Transactions on Image Processing*, 27(6):2609–2622, 2018. 1