Human Motion Segmentation via Velocity-Sensitive Dual-Side Auto-Encoder

Yue Bai, Lichen Wang, Yunyu Liu, Yu Yin, Hang Di, and Yun Fu, *Fellow, IEEE*

***Abstract*—Human motion segmentation (HMS) aims to segment a long human action video into a bunch of short and meaningful action clips. Existing supervised learning approaches require a considerable amount of data for well-training a model, which may be costly in real-world scenarios. Most unsupervised clustering methods cannot fully explore the temporal correlations among human motions and it is hard to achieve promising performances. In this work, we design a novel Velocity-Sensitive Dual-Side Auto-Encoder (VSDA) for HMS tasks. Specifically, a multi-neighbor auto-encoder (MNA) is proposed to extract informative temporal features, which fully explores the local temporal patterns of human motions. In addition, a long-short distance encoding (LSE) mechanism is designed. LSE constrains the encoded representations of close (short-distance) frames becoming similar while the representations of far-away (long-distance) frames becoming distinctive. Similarly, this strategy is also deployed on the decoded outputs as the long-short distance decoding (LSD) module. The proposed LSE and LSD guide the learning process explicitly and implicitly to achieve the dual-side structure. Moreover, we consider the energy variations in the human motion data, and propose the velocity-sensitive (VS) guidance mechanism for further model improvement. VSDA leverages the temporal characteristics of human motion and derives promising HMS performance. Extensive experiments and ablation studies demonstrate the effectiveness of our VSDA model.**

***Index Terms*—human motion, unsupervised learning, auto-encoder, energy-based**

I. Introduction

Human motion segmentation (HMS) divides a long action video into several short clips. Each clip has its own meaning. The basic concept is shown in Fig. 1. For instance, one real-world video usually contains hundreds of clips, most existing classification or recognition methods [1] are designed for dealing with clips containing only one single action. Therefore, HMS methods are required for cutting the raw video into several short clips for the down-streaming tasks. HMS is an indispensable data pre-processing step for many motion/action related tasks (e.g., motion analysis, action recognition, and security surveillance).

HMS is a challenging task due to the complicated temporal characteristics among the high-dimensional motion features even in multiple views [2]–[5]. It mainly focuses on exploring efficient and effective clustering-based approaches to gather correlated motion frames to segment in the unsupervised learning scenario. Different with the static data, the dynamic temporal correlations play a critical role for segmentation tasks [6]. As a summary, there are two main difficulties for handling HMS [6]. First, human motion data contains complicated frame-level temporal correlations. Second, human motion data contains complex dynamic patterns existing in the whole motion sequence.

Effectively modeling successive temporal information is the key factor for HMS. Based on the strategies of modeling temporal information, existing motion analytical approaches can be grouped in three categories [7]: 1) representation learning based [8]–[10], 2) model based [11], and 3) temporal proximity based [6]. Most existing HMS methods belong to representation learning based strategy. They firstly derive distinctive representations from original video samples. Then, these representations are set as input for down-streaming clustering algorithms (e.g., K-means [12] and Normalized Cuts [13]) to obtain final segmentation results. To name a few, a temporal constraint is utilized to obtain representations for performance improvement [14]. Specific dictionaries are designed to derive distinctive representations [15], [16]. Highly correlated frames are considered to enhance the feature capacities in [8], [17]. [18] proposed a graph based embedding strategy. It updates graph representation dynamically and serves for first-person video segmentation which is another important application for temporal clustering. [19] introduced Dynamic Graph Embedding (DGE) for event representation learning. It jointly learns the graph and the graph embedding via an iterative optimization strategy in the unsupervised manner. [20] further explores the dependencies in long-range for more distinguished frame representation learning.

Transfer learning strategy utilizes existing human motion dataset as auxiliary or source information to guide the representation learning on target data [21]. [22] introduced a multi-mutual induced learning method. It extracts the frame-level features from multiple network layers which achieves high and stable performance. [23] further explores the diversity and consistency of human related motion signals.

However, these algorithms have several drawbacks. First, most of them use traditional optimization algorithms which need high computational resources (e.g., eigen decomposition). Second, most approaches globally model the human motion information while ignoring the trivial local details among the temporal domain. Third, few existing algorithms consider the temporal dynamic characteristic (e.g., motion energy variations) to comprehensively guide the representation learning process. As a result, these approaches are easy to

Yue Bai, Lichen Wang, Yunyu Liu, Yu Yin, Hang Di are with the Department of Electrical and Computer Engineering, Northeastern University, Boston, USA (Email: {bai.yue, wang.lich, liu.yuny, yin.yu1, di.h}@northeastern.edu).

Yun Fu is with the Department of Electrical and Computer Engineering, and Khoury College of Computer Science, Northeastern University, Boston, USA (Email: yunfu@ece.neu.edu).

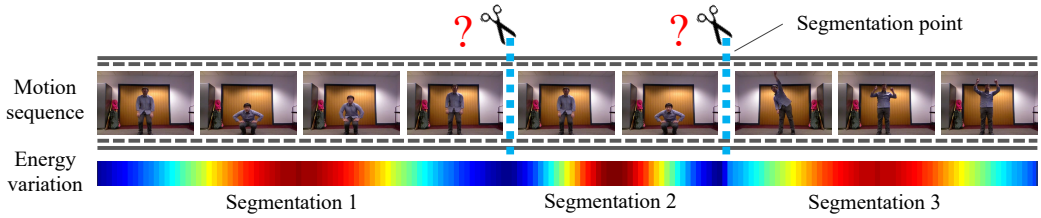


Fig. 1. Human motion segmentation (HMS) focuses on segmenting a long motion video into many short and meaningful action clips (e.g., running, jumping, and walking). Obtaining robust representations of temporal input is important to achieve accurate results, especially near the boundary of two adjacent clips. The color bar represents the energy variations which is an important clue to facilitate temporal segmentation. Warmer color represents higher energy variation during action; colder color means relatively low energy variation.

derive inaccurate segmentation results especially in action-switch regions. They not only diminish the segmentation performance, but also exert negative impacts on down-streaming tasks and limits the potential usage of the algorithm in real-world applications.

We designed a Velocity-Sensitive Dual-Side Auto-Encoder (VSDA) framework to handle the above challenges for HMS. Fig. 2 illustrates the framework of VSDA. A specifically designed multi-neighbor auto-encoder (MNA) is first proposed, where the input is a single frame and the reconstruction targets are multiple neighbor frames of the input. By this way, the local structural information will be preserved in the encoded representations. In addition, a long-short distance constraint is proposed. The insight is that the close frames (short-distance) in time domain are highly possible belonging to the same action, while the far-away (long-distance) frames are supposed to be in different actions. To this end, the long-short distance encoding (LSE) is utilized to pull the short-distance representations together while separating the long-distance representations. Similarly, the long-short distance decoding (LSD) is further deployed on the decoded side. The LSE and LSD guide the learning process explicitly and implicitly to achieve a dual-side structure. Moreover, we consider the energy variations during the human motions to fully explore its dynamic characteristics. We assume the motion has high velocity (energy variation) in the middle, while it has low velocity in the beginning and ending parts. The high velocity in the middle easily causes the inconsistent representations within the motion, which should be clustered into the same segment. These "velocity-sensitive" frames are crucial to improve the model performance. Thus, we should pay more attention to the frames with high energy variations. Specifically, we calculate the energy variations to capture dynamic patterns as velocity-sensitive (VS) guidance. Then, we utilize it to adaptively adjust the proposed long-short distance constraint. To summarize, our main contributions are list below:

- A multi-neighbor auto-encoder (MNA) framework is designed. It captures the action information associated with the neighbor frames to preserve the local temporal patterns.
- The long-short distance objectives are explored on encoding (LSE) and decoding (LSD) stages to achieve a dual-side structure. It obtains stable representations of neighbor frames while differentiating the representations of the long-distance frames. The LSE and LSD guide the model explicitly and implicitly to obtain robust representations.
- The human motion energy variation is considered as velocity-sensitive (VS) guidance. Based on our previous auto-encoder framework, the VS module guides the model focus on sensitive frames with high velocity. In this way, the long-short distance constraint is adaptively adjusted to achieve better segmentation results.

There are several advantages of our VSDA: 1) Both the local (short-distance) and global (long-distance) temporal correlations are well preserved and explored for high quality representation learning; 2) The motion energy factor is fully explored to globally guide the representation learning, which considers the temporal dynamic characteristics to model human motion data; 3) All modules can be achieved concisely and efficiently, which will be suitable for large-scale practical applications without high computational cost. Extensive experiments are conducted to show the model effectiveness.

Compared with our previous version conference paper [24], we summarize the differences as follows: First, based on our dual auto-encoder structure, we further consider the velocity factor to fully explore the human motion data. It digs more temporal dynamics and can be easily plugged into our previous framework. In addition, we provide more model details as well as the experimental analysis (e.g., time consumption analysis, training process visualization, and extra ablation study experiments). Extensive experiments on the modified framework demonstrate the proposed VSDA obtains higher segmentation performance than our conference version. We organize the literature review in Sec. II, present our proposed method in Sec. III, experimentally evaluate our model in Sec. IV with relevant analysis, and draw the conclusions in Sec. V.

II. Related Work

Most existing methods for Human Motion Segmentation (HMS) utilize temporal data clustering techniques. Besides, our model is designed based on the auto-encoder structure. To this end, we mainly introduce temporal data clustering and auto-encoder related approaches for human motion analysis in this section.

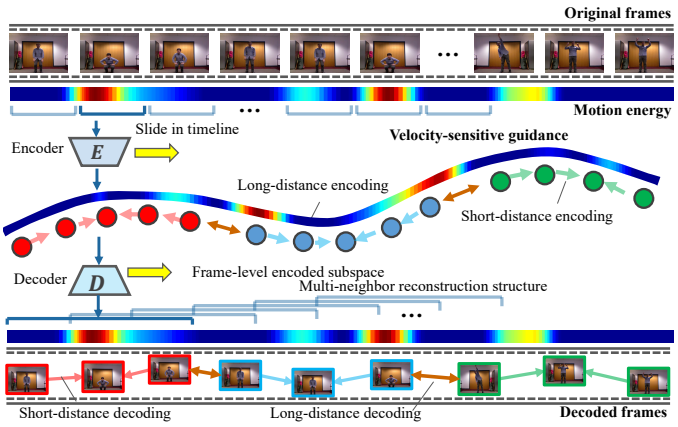


Fig. 2. The framework of our proposed VSDA approach. Frame-level features are extracted from the human motion samples. Then, the features are set as input to the multi-neighbor auto-encoder (MNA). The long-short encoding/decoding (LSE/LSD) strategies are applied as a dual-side structure to guide the representation learning process explicitly and implicitly. By this way, the similarity and distinctiveness between frames are learned and preserved in the representations. The motion velocity (energy) based space mapping weights are calculated to re-distribute the representation space and obtain the robust representations. Finally, the obtained representations are forwarded to a down-stream clustering algorithm (e.g., NCuts) for the final segmentation results.

A. Temporal Data Clustering

Temporal clustering algorithms aim to segment long temporal data into many short and rational clips. It can be widely utilized in applications including NLP, speaker diarization, and human motion segmentation. Hierarchical cluster analysis [25] designs a dynamic kernel to align the temporal features for clustering tasks. Semi-Markov K-means clustering [26] aims to capture temporal features in time series data. [27] proposed a Maximum-margin method which recognizes the position and length of the short segments. A specific dictionary and corresponding representations are jointly learned via a temporal subspace learning method with a regularization to extract temporal information [16]. A non-local self-similarity learning function is deployed to benefit temporal segmentation [20]. A dynamic graph embedding strategy is proposed to conduct temporal segmentation in a sequence of images [19]. A transfer learning mechanism is used to explore the human motion knowledge from extra motion videos to boost the segmentation performance [21]. A low-rank constraint is involved in the optimization procedure to achieve the better results [8], [28]. To further explore the knowledge transferring from source to target dataset, a subspace learning strategy in multi-level is designed in [22], [23]. They leverage more fine-grained information to improve the target performance. Graph data explores connections across different samples or nodes [29]. A novel graph-based constraint is proposed for motion segmentation in the unsupervised learning scenario [18]. Basically, these methods mainly use traditional optimization methods which have high computational cost. In addition, most of them globally model human motion patterns while ignoring the local information. Our proposed VSDA framework employs both global and local characteristics for learning valuable frame representation. Further, our VSDA involves the motion energy factor to adaptively adjust the learning process.

B. Auto-encoder for Human Motion

Auto-encoder structures designed for human motion analysis mainly focus on learning informative representations, which are usually set as input for down-streaming tasks such as classification and clustering. Auto-encoder based frameworks are used for several applications such as action recognition and human motion prediction. A deep auto-encoder is designed for pose recovery via multi-modal feature extraction [30]. A convolutional auto-encoder is designed to learn human motion manifolds in a unsupervised learning scenario [31]. Action forecasting can be handled by using a specific designed conditional variational auto-encoder to make generation [32]. In addition, an auto-encoder for denoising in stacked manner is wisely defined for action recognition [33]. However, these approaches use the auto-encoder structure directly without further enhancement. In this paper, we explore the temporal correlations with involving energy factor to develop the velocity-sensitive dual-side auto-encoder (VSDA) for HMS.

III. Methodology

A. Prelilminary

We first introduce the setting of human motion segmentation (HMS). Let $X \in \mathbb{R}^{T\times f}$ be the human motion input, where T is motion sequence frame numbers and f is the feature dimension. Our Velocity-Sensitive Dual-Side Auto-Encoder (VSDA) model aims to derive the valuable temporal representations of X for each t $(1 < t < T)$. Next, the down-streaming clustering algorithm, Normalized Cuts (NCuts) [13], is used to obtain the final segmentation results. VSDA contains four major components: 1) multi-neighbor auto-encoder (MNA) which is the basic structure of our complete framework, 2) long-short encoding (LSE) which is employed on the encoding process to explicitly direct the representation learning, 3) long-short decoding (LSD) which is symmetrically utilized on decoding to implicitly enhance the learning part, 4) the velocity-sensitive (VS) guidance which fully considers the energy variations of human motion and adjusts the proposed long-short constraint. The details of each component will be introduced below.

B. Multi-Neighbor Auto-Encoder

Auto-encoder is an effective and widely used structure to extract informative representations in the unsupervised learning scenario. It uses the encoding and decoding procedures to obtain low-dimensional and informative representations. However, conventional models focus on reconstructing the raw input but cannot consider and preserve the temporal local patterns for human motion sequence. The temporal local correlation is crucial to achieve high segmentation performance in HMS. Therefore, the learned representations should not only contain the input information but also preserve the temporal local patterns. To this end, we propose a MNA structure. Specifically, the input is the feature of each single frame, while the reconstruction targets are the multiple neighbor frames of

Algorithm 1 The pesudo code of training VSDA algorithm.

Require: Input data $\mathcal{X}$, number of training steps S
Ensure: Clustering results Y of motion sequence $\mathcal{X}$

1: **for** each $i \in [1, S]$ **do**
2: Encode $\mathcal{X}$ to obtain initial features X (e.g., obtaining HoG features in our experiments)
3: Forward X into MNA $E_e(\cdot)$ and $E_d(\cdot)$, and compute H_e and H_d through Eq. (1) and Eq. (2)
4: Add long-short distance constraints on both encoding and decoding sides by Eq. (7) and Eq. (10)
5: Compute the motion energy patterns by Eq. (11) and enhance the long-short distance constraint using VS guidance by Eq. (12)
6: Jointly update the whole framework by optimizing Eq. (13) to obtain H_e.
7: **end for**
8: Forward H_e into the down-streaming NCuts algorithm and obtain final segmentation results Y
9: **reture** Y

the original input. Based on this training strategy, the learned representation will contain the temporal local information for the segmentation. The MNA contains two modules. First is a single-frame encoder $E_e(\cdot)$ given by:

$$h_e^t = E_e(X^t), \tag{1}$$

and a multi-frame decoder $E_d(\cdot)$ given by:

$$h_d^t = E_d(h_e^t), \tag{2}$$

where $X^t \in \mathbb{R}^f$ is the t-th raw frame of X for $t \in \{1, 2, 3, ..., T\}$. h_e^t is the representation from a hidden layer, and h_d^t is the recover results of X^t. Both $E_e(\cdot)$ and $E_d(\cdot)$ are implemented by a linear mapping with ReLU activations. To achieve the multi-neighbor reconstruction, the input frame associated with its neighbor frames are assigned as the auto-encoder loss. The MNA is formulated by optimizing the following objective on t-th frame:

$$L_a^t = \sum_{k=t-w}^{t+w} \|h_d^t - X^k\|_{\mathrm{F}}^2, \tag{3}$$

where w represents the scale for neighbor reconstruction. $\|\cdot\|_{\mathrm{F}}^2$ is l_2-norms. u and v are two given frames. At the head and tail of the region of the given video, we explore the intermediate frames to be reconstruction targets. Following this path, the h_e^t is directed to restore local motion knowledge from its neighbor frames. The MNA objective function on the whole motion sequence is given by

$$L_a = \sum_{t=1}^{T} L_a^t. \tag{4}$$

We introduce our MNA by defining Eqs. (1)-(4). It is the basic structure of our VSDA and obtains the initial temporal features h_e^t to lay a solid foundation for the following model component. The other modules will be introduced based on the MNA.

C. Long-Short Encoding

The MNA aims to preserve local temporal information through a well-designed reconstruction strategy. However, HMS also relies on the feature correlations to accurately segment each motion frame. To this end, h_e^t are supposed to be as same as possible in the same action clip and vice versa. We propose a long-short encoding (LSE) approach to improve the distinctiveness of the learned representation. The insight is straightforward that the frame-level h_e^t which are close in the temporal domain (short-distance) should be similar while those which are far-away from each other motion (long-distance) should be different. The LSE contains two constraints. The first is short-distance constraint which is defined by the following objective:

$$L_l = \sum_{t=1}^{T} \sum_{k=t-s}^{t+s} \|h_e^t - h_e^k\|_{\mathrm{F}}^2, \tag{5}$$

where h_e^t denotes the outputs from a hidden layer, and h_e^k are the neighbors of h_e^t. s is the constraint length. L_l is the objective function of the short-distance constraint. This strategy enhances the local similarity. For example, if close frames belonging to the same segment have large-scale variations, they will be easily segmented into several fractions which results in low segmentation performance. Our short-distance constraint reduces the variety and smooths the learned representation to achieve better results.

On the other hand, in a long motion sequence containing several segments, frames temporally far-away from each other should be in different segments. The long-distance distinctive constraint makes these features more distinctive. It is defined by following objective:

$$L_g = \sum_{t=1}^{T} \Big(\sum_{k=t+q}^{t+q+s} \|h_e^t - h_e^k\|_{\mathrm{F}}^2 + \sum_{k=t-q-s}^{t-q} \|h_e^t - h_e^k\|_{\mathrm{F}}^2 \Big). \tag{6}$$

The long-distance constraints of t-th representation are summed together. q denotes the difference in long-distance perspective of the t-th and the target frames. The outer summation is the constraint of the whole motion sequence. The long-distance constraint lets representations of the far-away frames be more distinctive. It avoids the dispersive frames from different segments being clustered together. For example, some far-away frames may be similar such as "walking" and "running" both contain a "standing" motion. They are easily clustered together and damage the final performance.

The complete LSE is to maximize L_g and minimize L_l simultaneously, which can be formulated as minimizing following objective:

$$L_h = L_l - \theta_h L_g, \tag{7}$$

where θ_h is a trade-off parameter.

D. Long-Short Decoding

In LSE, we deploy constraints on hidden representations h_e^t. In the auto-encoder model, the h_e^t is used to reconstruct the raw input. Therefore, h_d^t distribution could also influence the learning results of h_e^t. To this end, the long-short constraint on

h_d^t is proposed which helps to stabilize the training procedure. The short-distance objective is below:

$$L_m = \sum_{t=1}^{T} \sum_{k=t-s}^{t+s} \|h_d^k - h_d^t\|_{\mathrm{F}}^2, \tag{8}$$

where h_d^k are the neighbors of h_d^t with the constraint distance s. In addition, the long-distance distinctive objectives are shown below:

$$L_n = \sum_{t=1}^{T} (\sum_{k=t+q}^{t+q+s} \|h_d^t - h_d^k\|_{\mathrm{F}}^2 + \sum_{k=t-q-s}^{t-q} \|h_d^t - h_d^k\|_{\mathrm{F}}^2). \tag{9}$$

LSD is realized by maximizing L_n and minimizing L_m which is shown below:

$$L_r = L_m - \theta_r L_n, \tag{10}$$

where θ_r is a trade-off parameter.

The long-short strategies are simultaneously optimized to achieve the dual-side structure. In this way, the representation learning is enhanced explicitly and implicitly.

E. Velocity-Sensitive Guidance

The MNA, LSE, and LSD modules comprehensively explore the temporal correlations in motion sequence. However, they treat each frame equally and ignore the different importance degrees of different frames. Specifically, human motion data always has high energy variations (velocity) in the middle, while has low variations in the beginning and ending. The middle frames belonging to the same motion with high velocity are more sensitive and unstable. They are easily clustered into different groups and damage the performance. These velocity-sensitive (VS) frames should be emphasized to enhance the representation learning. To this end, we take the motion velocity into account as guidance information for our model. We compute the motion energy variations based on velocity changes to capture the dynamic motion velocity patterns. The energy of each frame is given by

$$e^t = log(\frac{|X^t - X^{t-1}| + |X^{t+1} - X^t|}{2}), \tag{11}$$

where e^t is the energy vector derived for X^t based on its neighbors. We take the summation on feature dimension for e^t to obtain the scalar value E^t. The energy vector $E = \{E^1, ..., E^T\}$ is the guidance knowledge to indicate the importance degree of each frame. In this way, we enhance our proposed long-short distance constraint by

$$\hat{L_*^t} = E^t \cdot L_*^t, \tag{12}$$

where L_*^t is t-th objective of L_*, where $* \in \{l, g, m, n\}$. $\hat{L_*^t}$ is the enhanced objective. Correspondingly, we have enhanced $\hat{L_h}$ and $\hat{L_r}$ for Eq. (7) and Eq. (10), respectively. Leveraging on the proposed VS guidance, the auto-encoder architecture adaptively adjusts the constraint frame-by-frame to achieve higher segmentation performance.

Our complete VSDA consists of the MNA, LSE, LSD, and VS guidance modules. Each module can be described by L_a, $\hat{L_h}$, and $\hat{L_r}$, respectively. The complete VSDA is realized by optimizing the loss function shown below:

$$L = L_a + \lambda_h \hat{L_h} + \lambda_r \hat{L_r}, \tag{13}$$

where λ_r and λ_h are trade-off parameters. In summary, VSDA obtains informative structural information from the original high-dimensional motion sequence. The learned representations h_e^t, as the final features, are forwarded to the downstreaming NCuts clustering algorithm for performance evaluations.

F. Discussion

As our complete framework involves several components in total, we supplement some discussions here to summarize them. Our Velocity-Sensitive Auto-Encoder (VSAE) contains DASE [24] and the newly extended velocity guidance module. VSAE contains three main parts: multi-neighbor auto-encoder (MNA), long-short encoding, and decoding constraint. Concretely, the MNA regularizes the learning process during the reconstruction by enforcing the hidden representation construct multiple adjacent frames. The long-short encoding focuses on the hidden representations by pulling (pushing) local (non-local) frame features close (far-away) from each other. It guides the learning explicitly as the learned hidden representations will be used for downstream segmentation. On the other hand, the long-short decoding is designed for decoded representations with the similar functions above. It guides the learning implicitly since the decoded representations are not used for segmentation but will affect feature capacity of hidden representations. Please note that Eq. (3) is the reconstruction loss between the decoded representations and groundtruth. Eq. (8) is added on adjacent decoded representations. These two terms serve similar functions but from two different aspects. The extended velocity based module is an extra part to involve energy information to further guide the learning process for the basic DASE framework.

G. Clustering

After obtaining the representations h_e^t, we deploy a cluster algorithm to perform final segmentation. In this work, we follow the setting of [8] and utilize Normalized Cuts (NCuts) [13] clustering algorithm. Existing algorithms such as LRR [34] and SSC [15] regarded the clustering graph weights as $(|H_e| + |H_e^\top|)/2$. Nevertheless, the temporal data always contains highly correlated within-cluster samples [35]. To handle the challenge, we explore the distance metric of [16] to calculate the similarity. W_G is built based on the metric between the pairwise of the representations:

$$W_G(i,j) = \frac{H_{ei}^\top H_{ej}}{||H_{ei}||_2 ||H_{ej}||_2}. \tag{14}$$

When W_G is obtained, NCuts is used to get the segmentation results. We assume the number of clusters is given. We summarize our whole framework in Algorithm. 1.

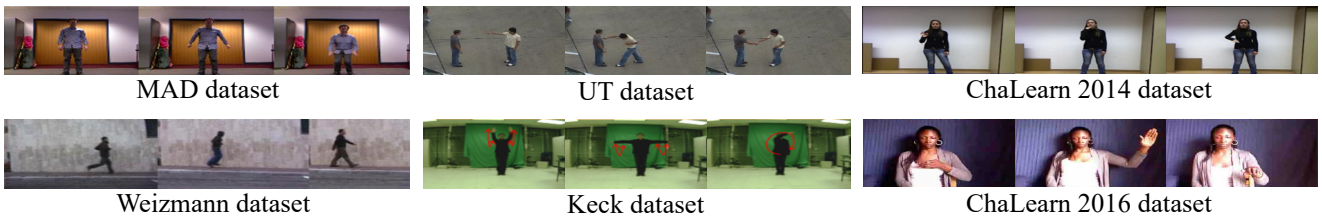


Fig. 3. The samples of six human motion datasets in our experiments. They are MAD, UT, ChaLearn 2014, Weizmann, Keck, and ChaLearn 2016, respectively, including single-subject and multi-subject interactive motions.

TABLE I
SEGMENTATION PERFORMANCE COMPARISON ON SIX DATASETS BASED ON ACC AND NMI METRICS.

Datasets	MAD		Keck		Weizmann		UT		ChaLearn16		ChaLearn14	
Methods	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KMD [36]	0.3226	0.3914	0.3970	0.4702	0.4441	0.5289	0.5122	0.5108	0.4160	0.2946	0.5078	0.6270
KMS [37]	0.3541	0.4188	0.3510	0.4553	0.4081	0.5562	0.4712	0.5677	0.4331	0.3221	0.4523	0.5968
LRR [34]	0.2397	0.2249	0.4297	0.4862	0.3638	0.4382	0.4162	0.4051	0.3239	0.1423	0.4137	0.5033
SPE [38]	0.3639	0.4369	0.3886	0.4744	0.4127	0.5435	0.4477	0.4894	0.4066	0.2721	0.4359	0.5877
SSC [15]	0.3817	0.4758	0.3137	0.3858	0.4576	0.6009	0.4389	0.4998	0.3867	0.2108	0.4853	0.6788
OSC [14]	0.4327	0.5589	0.4393	0.5931	0.5216	0.7047	0.5846	0.6877	0.4025	0.3346	0.4759	0.7189
LSR [17]	0.3979	0.3667	0.4894	0.4548	0.5091	0.5093	0.5183	0.4322	0.3917	0.1973	0.5913	0.5817
TSC [16]	0.5556	0.7721	0.4781	0.7129	0.6111	**0.8199**	0.5340	0.7593	0.5414	0.6000	0.5373	0.7861
TSS [21]	0.4652	0.6987	0.4929	0.7342	0.6101	0.7112	0.5541	0.7114	0.5385	0.6410	0.3788	0.6602
LTS [8]	0.4833	0.7268	0.5128	0.7365	0.6155	0.7273	0.5629	0.7223	0.5359	0.5369	0.3734	0.5684
DSAE [24]	0.5548	0.7734	0.5753	**0.7407**	0.6199	0.7879	0.6006	0.7950	0.5905	0.6673	0.6055	0.8515
VSDA(Ours)	**0.5606**	**0.7770**	**0.5804**	0.7397	**0.6287**	0.7992	**0.6203**	**0.8226**	**0.6007**	**0.6826**	**0.6291**	**0.8578**

IV. EXPERIMENTS

A. Datasets

In our experiments, six real-world human motion datasets are used to evaluate our model. They are collected in different environments and contain various human actions including single-subject action (e.g. "walking" and "running") and multi-subject interact actions (e.g. "hugging" and "punching"). We introduce the dataset details as below and show a few action samples in Fig. 3.

- **Multi-Modal Action Detection Dataset (MAD)** [39] has human motions performed by 20 subjects collected in RGB, depth, and skeleton modalities. All models are captured synchronously at 30 fps. The RGB frames are captured in 240×320 resolution with 35 actions performed by each subject. In our experiments, we only use the videos in RGB modality.
- **UT-Interaction Dataset (UT)** [40] has 6 types of motions including handshaking, pointing, pushing, hugging, kicking, and punching. Each video lasts around 1 minute.
- **Weizmann Dataset (Weiz)** [41] includes 90 motion samples including 10 motions such as running, walking, and skipping. It is collected from nine subjects and each subject performs each motion one time.
- **Keck Gesture Dataset (Kect)** [42] contains 14 different gestures and motions from 3 subjects performing military signals. It is collected by a fixed camera and the human subjects perform motions in front of a simple and static background. The RGB frames are collected with 640×480 resolution.
- **ChaLearn 2014** [43] contains 14,000 gestures about the vocabulary from 20 Italian sign gesture classes. It aims to perform user independent continuous gesture spotting. They are performed by several different users and used to evaluate our model.
- **ChaLearn 2016** [44] is a gesture dataset which has 47,933 gestures from 22,535 RGB-D motion recordings. Overall, 249 gestures categories are conducted by 21 human volunteers. It can be used for segmenting and recognizing gestures from a continuous video. RGB videos are used in our experiments.

B. Comparison Approaches

We compare our model with several competitive approaches to illustrate the effectiveness of our approach. They include some conventional clustering algorithms and several recently proposed SOTA models for HMS. The details of comparison methods as listed below.

- **K-means (KMS)** [37] aims to cluster each frame-level feature using the nearest mean which minimizes the sum of squres for within-cluster.
- **K-medoids (KMD)** [36] chooses targets as centers and clusters samples with a distance between points defined by a generalization of Manhattan Norm.

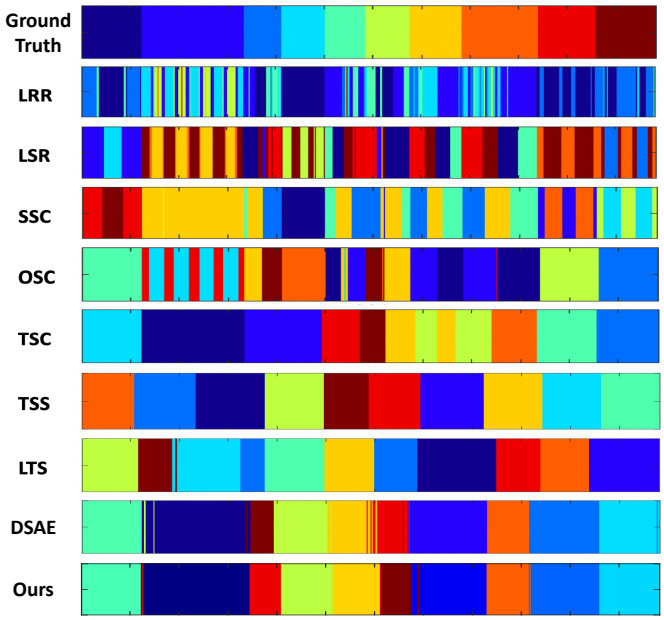


Fig. 4. We visualize the segmentation results based on the representation from different methods compared with ground truth. Different colors represent different segments. First four methods have low accurate results with many disorders, which is unacceptable for HMS tasks. Next four approaches achieve much better results but still accompanied with fractions in each segment and inaccurate segment boundaries. Ours obtains the best results without too many fractions in each segment and more accurate boundaries.

- **Spectral Clustering (SPE)** [38] employs the similarity matrix spectrum from the target samples to achieve dimension reduction and obtain better clustering results.
- **Low-Rank Representation (LRR)** [34] effectively obtains the sample global structure. It can deliver robust segmentation results from corrupted data which contains high-level outlier samples. The global structure of the samples can be derived to obtain more robust segmentation results.
- **Ordered Subspace Clustering (OSC)** [14] proposes an objective in temporal prespective during the learning process which directly lets the temporal representations be aligned in feature space.
- **Sparse Subspace Clustering (SSC)** [15] hypothesis there is a potential dictionary of each data sample. It proposes a sparse constraint to derive the coefficients of the dictionary and learns the sparse features of frames.
- **Least Square Regression (LSR)** [17] groups highly correlated data samples together by encouraging a specific grouping effect based on the Frobenius norm.
- **Temporal Subspace Clustering (TSC)** [16] explores a learnable dictionary and temporal Laplacian regularization to simultaneously get the informative representations for motion signals.
- **Transfer Subspace Segmentation (TSS)** [21] develops an approach based on transfer learning technique to obtain knowledge from betweem source and target samples. The auxiliary data information is leveraged to increase the results.
- **Low-Rank Transfer Segmentation (LTS)** [8] designs a graph model for sequential signals in a transfer learning fashion. A low-rank objective is proposed for higher segmentation accuracy.
- **Dual-Side Auto-Encoder (DSAE)** [24] designs a novel auto-encoder structure which comprehensively considers the temporal correlations to enhance the representation learning.

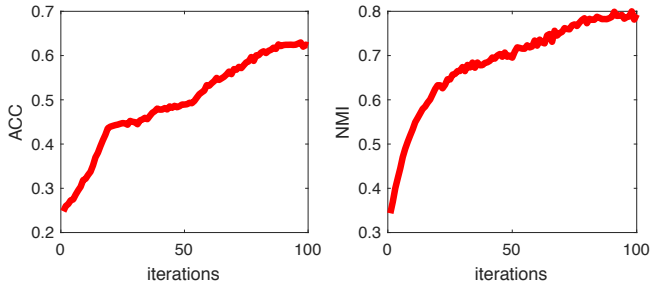


Fig. 5. The update trends of ACC and NMI for one sample over the iterations of model training. We find the model consistently and stably converged based on our proposed training strategy.

C. Implementation

HoG feature [45] is popular and efficient for representing human motion information which can be conducted conveniently to obtain frame-level features for the whole motion sequence. For all six datasets, we utilize the HoG encoding to derive the initial feature sequence with 324-dimension as VSDA model input. To avoid the inconvenience among different datasets, we standardize the input data for different datasets. Concretely, both Keck and Weizmann datasets contain a single motion in each video, we concatenate 10 single-action videos to obtain a long video following the same setting in [27]. The videos in the MAD dataset have more than ten actions. We remove the extra and reserve the first ten in each recording. For the UT dataset, there are only six actions in each recording. In this scenario, we keep it as it is. For the ChaLearn2014 dataset, we choose videos containing ten or more actions and reserve them into ten-action videos. For the ChaLearn2016 dataset, we choose all the video samples with more than 5 actions for consistent comparison.

The first three comparison methods, KMS, KMD, and SPE, are conventional clustering algorithms. The next four approaches, LRR, OSC, SSC, and LSR, are recently proposed segmentation models. TSC is specifically designed for HMS. All of them follow the common clustering experimental setting and we directly evaluate their learning performance. TSS and LTS are two HMS methods in the transfer learning scenario. They aim to use auxiliary knowledge from a source dataset to benefit the segmentation in target. Due to the setting difference, we unify the experimental setting for TSS and LTS to make a fair evaluation. Specifically, instead of involving another dataset, we set the target and source as the same to conduct their algorithm. In this way, the knowledge is derived from the target dataset itself instead of another dataset. DSAE follows conventional unsupervised setting, thus, we can directly make comparison.

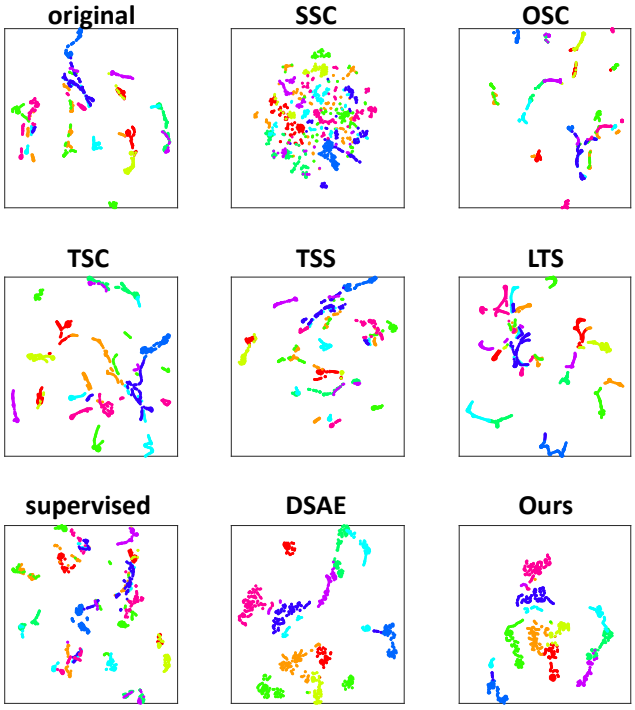


Fig. 6. We visualize the t-SNE of the embeddings from the original data, supervised learning model, and different unsupervised approaches. Compared with others, the embeddings from our method are well clustered, which demonstrates our model effectiveness.

D. Performance Analysis

We show qualitative analysis with more intuition to deliver a comprehensive evaluation for this practical task.

1) Numerical Evaluation: We use two numerical evaluation metrics in our work. Normalized Mutual Information (NMI) and Accuracy (ACC) [46]. They are formalized by using following equations:

$$ACC = \sum_{i=1}^{n} \frac{\delta(s_i \cdot map(r_i))}{n}, \tag{15}$$

$$NMI = \frac{\sum_{i,j} n_{ij} \log \frac{n n_{ij}}{n_{i+} n_{+j}}}{\sqrt{(\sum_j n_{j+} \log \frac{n_{+j}}{n})(\sum_i n_{i+} \log \frac{n_{i+}}{n})}}, \tag{16}$$

where $map(r_i)$ means the permutation mapping. The higher the value of NMI and ACC the better the performance.

The segmentation performances are illustrated on Table I. First three conventional clustering algorithms perform low in both ACC and NMI, since these strategies are not designed for handling human motion signals. The next four recently proposed approaches are representation based clustering algorithms. They still cannot obtain the best results. The TSC algorithm is specifically proposed for HMS and has promising results. Our model outperforms it in most scenarios. TSS and LTS are transfer learning based approaches. Nevertheless, if we adopt the evaluate setup for fair comparisons with our model, they cannot achieve better results. The detailed setting is introduced in the *Implementation* section. VSDA fully considers the temporal correlative patterns and achieves promising results, however, it ignores the dynamic motion characteristics. Our VSDA explores the temporal correlations and dynamic motion patterns simultaneously to obtain the best performance.

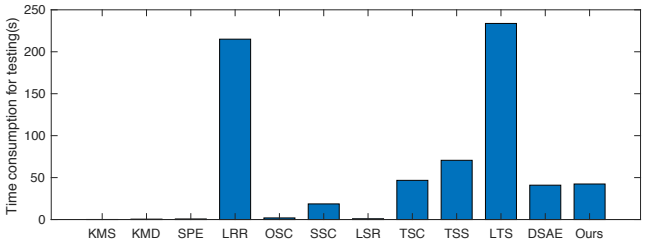


Fig. 7. Time consumption comparison. We find our method is faster than some other comparison approaches. Compared with several high-speed algorithms (e.g., KMS, KMD), ours achieves much better performance.

In Fig. 5, we visualize the ACC and NMI variations over model training iterations for one sample as shown. We notice that both NMI and ACC increase as the number of iterations goes up which means our proposed model works well during the training process. Further, near 100 iterations, the performances tend to be stable. These observations demonstrate that our proposed model is robust and effective to improve the clustering performance. Another important measurement of the HMS model capacity is time consumption.

2) Time Consumption: In Fig. 7, we compare the time consumption of the representation learning phase based on both our approach and the other approaches on the MAD dataset. We calculate the total processing time for 40 motion sequences in MAD. We observe that KMS, KMD, SPE, OSC, and LSR achieve very fast clustering, however, their segmentation performances are low. Our approach is more efficient than several competitors like TSS, LTS, and TSC. These methods require the iterative optimization strategies associated with computational costly steps such as eigen-decomposition. This leads to the inefficiency optimization procedure. Although our approach also needs to optimize the weights of the auto-encoder, the efficient gradient descent associated with parallel computing could significantly improve the speed of our model. Further, the conventional approaches and optimization methods will suffer from higher dimensional motion features. However, our method can be utilized efficiently on high-dimensional data with controllable time consumption. Please note, since our work focuses on temporal representation learning instead of the downstream segmentation, our time consumption comparison with other methods only considers the representation learning step. Both our model and other methods use the same downstream clustering algorithm for fair experiments.

3) Comparison with Supervised Learning: Our model is executed in the unsupervised learning scenario without frame-level label information. To show our model effectiveness, we further compare our VSDA with a basic supervised model. Specifically, we use the feature of each frame as training data with the corresponding action category as label. We train a simple linear classifier. Then, we extract features of the last hidden layer for NCuts to obtain segmentation results. We follow the leave-one-subject-out strategy to evaluate the

TABLE II
ABLATION STUDY ON HUMAN MOTION DATASETS

MNA	Short-En	Long-En	Short-De	Long-De	VS	Chalearn16		MAD		Kect		Weizman		UT		Chalearn14	
						ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
×	✓	✓	✓	✓	✓	0.5664	0.6537	0.5249	0.7263	0.5304	0.7199	0.5749	0.7298	0.5636	0.7660	0.5899	0.7802
✓	×	✓	✓	✓	✓	0.2637	0.2784	0.2368	0.3305	0.1910	0.2246	0.3210	0.4005	0.4911	0.5382	0.5013	0.6275
✓	✓	×	✓	✓	✓	0.5192	0.6409	0.4901	0.7427	0.4825	0.7306	0.5463	0.7407	0.5738	0.7826	0.5579	0.7927
✓	✓	✓	×	✓	✓	0.5751	0.6585	0.5388	0.7605	0.5714	0.7365	0.6142	0.7803	0.6112	0.8048	0.5812	0.8241
✓	✓	✓	✓	×	✓	0.5703	0.6498	0.5459	0.7598	0.5733	0.7428	0.6087	0.7717	0.6095	0.7994	0.5936	0.8241
✓	✓	✓	✓	✓	×	0.5905	0.6673	0.5548	0.7734	0.5753	0.7407	0.6199	0.7879	0.6006	0.7950	0.6055	0.8515
✓	×	×	✓	✓	✓	0.3832	0.5010	0.3338	0.4870	0.3580	0.4836	0.3547	0.5823	0.4697	0.5239	0.4792	0.6033
✓	✓	✓	×	×	✓	0.5524	0.6592	0.5234	0.7437	0.5322	0.7538	0.5610	0.7269	0.5597	0.7423	0.5234	0.7786
✓	×	✓	×	✓	✓	0.2419	0.2535	0.2987	0.3042	0.2550	0.2479	0.1945	0.2495	0.2909	0.2761	0.3081	0.3579
✓	✓	×	✓	×	✓	0.5058	0.6427	0.4981	0.7346	0.4900	0.7361	0.5324	0.7019	0.5804	0.7735	0.5661	0.7980
✓	✓	✓	✓	✓	✓	**0.6007**	**0.6826**	**0.5606**	**0.7770**	**0.5804**	**0.7397**	**0.6287**	**0.7992**	**0.6203**	**0.8226**	**0.6291**	**0.8578**

TABLE III
SUPERVISED METHOD (LINEAR CLASSIFIER) V. VSDA

Methods	Supervison	Keck		MAD		Weiz	
		ACC	NMI	ACC	NMI	ACC	NMI
Classifier	Yes	0.3855	0.4349	0.3976	0.4835	0.5081	0.5942
VSDA	No	**0.5804**	**0.7397**	**0.5606**	**0.7770**	**0.6287**	**0.7992**

models on three datasets. Table. III shows their performance comparisons. We conclude the supervised method cannot outperform our proposed model. Our VSDA has better results in the unsupervised scenario without high labeling costs.

4) Segmentation Visualization: The segmentation results are illustrated in Fig. 4. Various colors means various motion clips. LSR and LRR results are low since there are fragments among the whole video since they ignore the temporal similarity and distinctiveness between different frames belonging to the same or different actions. Time steps belonging to the same segment are easily separated. Compared with the groundtruth label, it is hard to recognize the boundary of each cluster and unacceptable for HMS tasks. SSC and OSC are able to obtain more rational results for partial sequences with clearer segments boundaries. However, for those motion signals with the rhythmed patterns (e.g. repetitive motion), there are still many fragments in their results. Because of ignoring the temporal connections, the overall results of these two methods are still unsatisfactory. TSC has better performance with clear cluster boundaries and more accurate segments. However, there are still many fragments and the boundaries are blurry. LTS and TSS have better results with less redundant fragments, but they always make mistakes about recognizing boundaries. DSAE performs promising results. Our current model takes the energy variation into account and obtains the best segments. Specifically, compared with ground truth, our model obtains a more accurate segmentation boundary than DSAE.

5) t-SNE Visualization: We visualize the learned representations by utilizing t-SNE [47] approach in Fig. 6. The visualized sample is picked from the Keck dataset and different colors represent different clusters. The visualization includes the original data, the representation from five competitive approaches, the supervised method, and our VSDA. Compared with other approaches, our embeddings are well clustered without much disorder. It is worth to note that even if our model achieves promising results yet we can still find some inconsistencies in the t-SNE plot (e.g., in "Ours", the features in light blue are not clustered perfectly with some separations). These separated features are most likely caused by some inaccurate clustering around the boundary of adjacent motion clips. Like shown in Fig. 4, our method has generally decent results but with few inaccurate fragments. However, compared with other methods, ours generally obtains the most reasonable segmentation results.

E. Ablation Study

Our proposed VSDA consists of MNA, LSE, LSD, and VS guidance modules. Ablation studies we explored here is to demonstrate the usefulness of the proposed modules. The ablated model results for all datasets are shown in Table. II. Our integrated framework can be separated into different partitions based on both single module and module combinations. Specifically, for each column, "MNA", "En", "De", "Short", "Long", and "VS" represent multi-neighbor auto-encoder, encoding, decoding, short-distance constraint, long-distance constraint, and velocity-sensitive guidance, respectively. For all the ablation tables, the upper block contains ablated models with removing each single module. The middle block contains ablated models with removing module combinations simultaneously. The lower block is our complete framework.

According to the ablation results from all tables, we make conclusions as: 1) For the upper block (removing single module), encoding constraint is more important than decoding constraint as the encoding constraint regularizes the model explicitly while decoding constraint is in an implicit way. The performance drops a lot for both ACC and NMI, especially short-distance encoding. However, the decoding constraint also improves the model performances which is also a necessary component. On the other hand, we can find the short-distance constraint provides a reliable foundation for representation learning, while the long-distance constraint further increases the model capacity. 2) For the middle block (removing module combinations), the results show that encoding and short constraint are generally more important than decoding and long constraint. This observation is consistent with that in the upper block. 3) We conclude the proposed constraint should

be used in a pairwise fashion and they can jointly improve the learning performance. 4) The ablated model by removing "MNA" is shown in the first row, which demonstrates the necessity of our multi-neighbor auto-encoder. 5) Compared with our basic version in the last row of upper block, the newly proposed velocity-sensitive component further boosts the final performance and our complete model achieves the highest segmentation results.

F. Future Work Discussion

The transformer network architecture dominants the deep learning field based on powerful self-attention mechanism. It globally models the dependency of the sequential data which is also expected to achieve higher learning performance for temporal segmentation task. Since this work mainly aims to involve a velocity module to enhance our previous auto-encoder structure, we leave the exploration of transformer model in our future work. In addition, the effective hand-crafted HoG feature is used in our work. This feature is extracted by considering the visual characteristic based on the frame image itself. Nowadays, using pretrained large-scale deep model to extract feature obtains promising performance. However, in this way, the extracted features are always relevant to semantic information. It may not help for the temporal segmentation which requires more internal patterns of a given temporal sequence. How to effectively employ pretrained deep model is another valuable point to explore in our future work.

V. Conclusions

We design a Velocity-Sensitive Dual-Side Auto-Encoder (VSDA) model for human motion segmentation (HMS) in an unsupervised scenario. A multi-neighbor auto-encoder (MNA) is employed for getting local structural knowledge. A long-short encoding (LSE) strategy is utilized on the learned representations to leverage temporal correlations explicitly. Similarly, the long-short decoding (LSD) is symmetrically employed on the decoding part to guide the model implicitly. The novel proposed velocity-sensitive (VS) guidance mechanism is used to enhance the dual-side constraint for further model improvement. Experiments based on six real-world human motion datasets show the effectiveness of our VSDA. A comprehensive ablation study proves each component in our model is effective and indispensable for achieving the highest performance. Besides, we also provide several model analyses including the segmentation visualization, t-SNE plot, and the time consumption comparison.

References

[1] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2015, pp. 4305–4314.

[2] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 333–342.

[3] J. M. Kleinberg, "An impossibility theorem for clustering," in *Advances in neural information processing systems*, 2003, pp. 463–470.

[4] Y. Liu, L. Wang, Y. Bai, C. Qin, Z. Ding, and Y. Fu, "Generative view-correlation adaptation for semi-supervised multi-view learning," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 318–334.

[5] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3413–3423.

[6] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: a survey and empirical demonstration," vol. 7, no. 4. Springer, 2003, pp. 349–371.

[7] Y. Yang and K. Chen, "Temporal data clustering via weighted clustering ensemble with different representations," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 2, pp. 307–320, 2010.

[8] L. Wang, Z. Ding, and Y. Fu, "Low-rank transfer human motion segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 1023–1034, 2019.

[9] Y. Bai, L. Wang, Z. Tao, S. Li, and Y. Fu, "Correlative channel-aware fusion for multi-view time series classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6714–6722.

[10] Y. Bai, Z. Tao, L. Wang, S. Li, Y. Yin, and Y. Fu, "Collaborative attention mechanism for multi-modal time series classification," in *Proceedings of the SIAM International Conference on Data Mining*, 2022, pp. 495–503.

[11] P. Smyth, "Probabilistic model-based clustering of multivariate and sequential data," in *Proceedings of the International Workshop on AI and Statistics*. San Francisco, CA: Morgan Kaufman, 1999, pp. 299–304.

[12] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[13] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Departmental Papers (CIS)*, p. 107, 2000.

[14] S. Tierney, J. Gao, and Y. Guo, "Subspace clustering for sequential data," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2014, pp. 1019–1026.

[15] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2009, pp. 2790–2797.

[16] S. Li, K. Li, and Y. Fu, "Temporal subspace clustering for human motion segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4453–4461.

[17] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proceedings of the European Conference on Computer Vision*. Springer, 2012, pp. 347–360.

[18] M. Dimiccoli, L. Garrido, G. Rodriguez-Corominas, and H. Wendt, "Graph constrained data representation learning for human motion segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 1460–1469.

[19] M. Dimiccoli and H. . Wendt, "Learning event representations for temporal segmentation of image sequences by dynamic graph embedding," *IEEE Transactions on Image Processing*, vol. 30, pp. 1476–1486, 2020.

[20] M. Dimiccoli and H. Wendt, "Enhancing temporal segmentation by nonlocal self-similarity," in *Proceedings of the IEEE International Conference on Image Processing*, 2019, pp. 3681–3685.

[21] L. Wang, Z. Ding, and Y. Fu, "Learning transferable subspace for human motion segmentation," in *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2018.

[22] T. Zhou, H. Fu, C. Gong, J. Shen, L. Shao, and F. Porikli, "Multi-mutual consistency induced transfer subspace learning for human motion segmentation," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2020, pp. 10 277–10 286.

[23] T. Zhou, H. Fu, C. Gong, L. Shao, F. Porikli, H. Ling, and J. Shen, "Consistency and diversity induced human motion segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[24] Y. Bai, L. Wang, Y. Liu, Y. Yin, and Y. Fu, "Dual-side auto-encoder for high-dimensional time series segmentation," in *Proceedings of the IEEE International Conference on Data Mining*, 2020.

[25] F. Zhou, F. De la Torre, and J. K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 582–596, 2012.

[26] M. W. Robards and P. Sunehag, "Semi-markov kmeans clustering and activity recognition from body-worn sensors," in *Proceedings of the IEEE International Conference on Data Mining*, 2009, pp. 438–446.

[27] M. Hoai and F. De la Torre, "Maximum margin temporal clustering," in *Artificial Intelligence and Statistics*, 2012, pp. 520–528.

[28] Q. Ma and A. Olshevsky, "Adversarial crowdsourcing through robust rank-one matrix completion," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, pp. 21 841–21 852, 2020.

[29] L. Wang, B. Zong, Q. Ma, W. Cheng, J. Ni, W. Yu, Y. Liu, D. Song, H. Chen, and Y. Fu, "Inductive and unsupervised representation learning on graph structured objects," in *Proceedings of the International Conference on Learning Representations*, 2019.

[30] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5659–5670, 2015.

[31] D. Holden, J. Saito, T. Komura, and T. Joyce, "Learning motion manifolds with convolutional autoencoders," in *SIGGRAPH Asia 2015 Technical Briefs*. ACM, 2015, p. 18.

[32] J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future: Forecasting from static images using variational autoencoders," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 835–851.

[33] A. Budiman, M. I. Fanany, and C. Basaruddin, "Stacked denoising autoencoder for feature representation learning in pose-based action recognition," in *IEEE Global Conference on Consumer Electronics*, 2014, pp. 684–688.

[34] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2012.

[35] S. Li and Y. Fu, "Learning balanced and unbalanced graphs via low-rank coding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1274–1287, 2014.

[36] L. K. P. J. RDUSSEEUN, "Clustering by means of medoids," in *Proceedings of the statistical data analysis based on the L1 norm conference*, 1987.

[37] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[38] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[39] D. Huang, S. Yao, Y. Wang, and F. De La Torre, "Sequential max-margin event detectors," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 410–424.

[40] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: video structure comparison for recognition of complex human activities." in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, 2009, p. 2.

[41] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, 2007.

[42] Z. Jiang, Z. Lin, and L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 533–547, 2012.

[43] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 459–473.

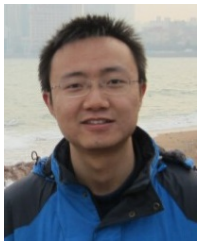
[44] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition," in *Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops*, 2016, pp. 56–64.

[45] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2005.

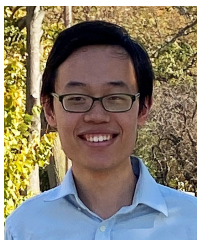
[46] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 877–886.

[47] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

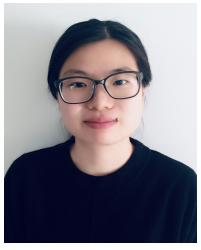
Yue Bai Received the B.Sc. degree in Mathematics from Donghua University, Shanghai, China in 2017. And the M.Eng. degree in Data Analytics Engineering from Northeastern University, Boston, MA, USA in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. His research interests include computer vision, machine learning, and deep learning.



Lichen Wang (S'16) received the B.Eng. degree in automation from Harbin Institute of Technology, Harbin, China in 2013, and the M.Eng. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China in 2016, and the Ph.D. degree in electrical and computer engineering from Northeastern University, Boston, USA in 2021. His research interests include machine learning, computer vision, natural language processing, data mining, and reinforcement learning. He serves as reviewer of many top-tier journals and conferences such as CVPR, ICCV, ECCV, ICML, ICLR, NeurIPS, TPAMI, TIP, TKDE, etc.

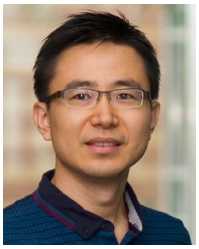


Yunyu Liu has received his B.Eng. degree in Information Engineering from Shanghai Jiao Tong University, China, in 2018 and M.S. degree in Electrical and Computer Engineering from Northeastern University, Boston, MA, USA, in 2020. He is currently pursuing the Ph.D. degree in the Computer Science Department at Purdue University. His research interests include multiview learning, deep learning and graph neural network. He has some top-tier conference papers accepted at ICCV, AAAI, ECCV et al.



Yu Yin Received B.S. in Electrical and Information Engineering (2016) from Wuhan University of Technology, China, M.S. in Electrical and Computer Engineering (2018) at Northeastern University (NEU), Boston, MA, and is currently pursuing a Ph.D. in Computer Engineering at NEU under Dr. Yun Fu. Her research spans image processing (*i.e.*super-resolution, face generation), visual recognition (*i.e.*face recognition, pose estimation, face alignment, and emotion recognition), and biosignal processing. She served as organizing chair of RFIW Workshop @ FG'20 and AMFG @ CVPR21, conference PC (*e.g.*AAAI & IJCAI & FG), and journal reviewer (*e.g.*IEEE Trans. on TIP & TNNLS).

Hang Di has received his B.Eng. degree in Electrical Engineering And Automation from Xi'an Jiao Tong University, China, in 2018 and M.S. degree in Electrical and Computer Engineering from Northeastern University, Boston, MA, USA, in 2020.



Yun Fu (S'07-M'08-SM'11-F'19) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary faculty member affiliated with College of Engineering and the Khoury College of Computer Sciences at Northeastern University since 2012. His research interests are Machine Learning, Computational Intelligence, Big Data Mining, Computer Vision, Pattern Recognition, and Cyber-Physical Systems. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; twelve Best Paper Awards from IEEE, ACM, IAPR, SPIE, SIAM; many major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an Associate Editor of the IEEE Transactions on Image Processing (TIP). He is Member of Academia Europaea, fellow of IEEE, IAPR, OSA and SPIE, a Lifetime Distinguished Member of ACM, Lifetime Senior Member of AAAI and Institute of Mathematical Statistics, member of ACM Future of Computing Academy, Global Young Academy, AAAS, INNS and Beckman Graduate Fellow during 2007-2008.