Frame-Level Label Refinement for Skeleton-Based Weakly-Supervised Action Recognition

Qing Yu,[1] **Kent Fujiwara**[2]
[1]The University of Tokyo, Japan
[2]LINE Corporation, Japan
[1]yu@hal.t.u-tokyo.ac.jp, [2]kent.fujiwara@linecorp.com

Abstract

In recent years, skeleton-based action recognition has achieved remarkable performance in understanding human motion from sequences of skeleton data, which is an important medium for synthesizing realistic human movement in various applications. However, existing methods assume that each action clip is manually trimmed to contain one specific action, which requires a significant amount of effort for annotation. To solve this problem, we consider a novel problem of skeleton-based weakly-supervised temporal action localization (S-WTAL), where we need to recognize and localize human action segments in untrimmed skeleton videos given only the video-level labels. Although this task is challenging due to the sparsity of skeleton data and the lack of contextual clues from interaction with other objects and the environment, we present a frame-level label refinement framework based on a spatio-temporal graph convolutional network (ST-GCN) to overcome these difficulties. We use multiple instance learning (MIL) with video-level labels to generate the frame-level predictions. Inspired by advances in handling the noisy label problem, we introduce a label cleaning strategy of the frame-level pseudo labels to guide the learning process. The network parameters and the frame-level predictions are alternately updated to obtain the final results. We extensively evaluate the effectiveness of our learning approach on skeleton-based action recognition benchmarks. The state-of-the-art experimental results demonstrate that the proposed method can recognize and localize action segments of the skeleton data.

Introduction

With the rise of demand from applications such as virtual reality and gaming, 3D skeleton motion data has become an important medium for understanding and synthesizing realistic human motion. Although skeleton motion data require dedicated apparatus for acquisition, the processed data can be directly applied to various characters to perform human movements. Moreover, marker-less 3D human motion capture from images is starting to make the acquisition of skeleton motion data easier without complex camera arrays and studios (Hasler et al. 2009; Habermann et al. 2019, 2020).

Recently, along with the growing interest in skeleton data itself, various action recognition (AR) methods based on skeleton data have also attracted considerable attention. As these methods are more focused on the posture of humans, skeleton-based methods have strong generalization ability and adaptability to complicated backgrounds (Vemulapalli, Arrate, and Chellappa 2014; Du, Wang, and Wang 2015; Tang et al. 2018). Skeleton-based models benefit from a number of advantages, such as strong robustness to variations in position, scale, and viewpoint, by utilizing the 3D coordinate positions of multiple key body joints to perform AR (Zheng et al. 2018).

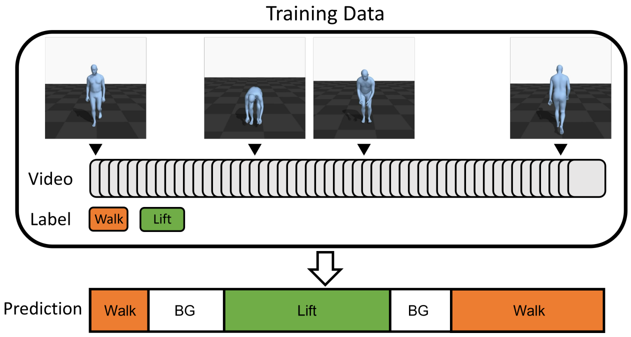

Figure 1: Illustration of S-WTAL. One example of the training data is a video stream of skeleton data, consisting of several action segments, *i.e.*, walk and lift, and transition frames denoted as background (BG). In the proposed S-WTAL task, the model needs to detect each action segment at frame-level from only the video-level labels. In other words, the frame-level ground truth is not available.

Conventional skeleton-based AR methods try to classify trimmed action clips, each containing a single action, into activity categories. However, acquiring such trimmed action clips requires an enormous amount of manual labor. This may not scale efficiently with the growing set of data size, total video length, and activity categories. The variance in the timestamps of each action (*i.e.*, the beginning and ending of each action) among annotators can also affect the quality of the dataset. On the other hand, it is much simpler for a human to offer a few labels that sum up the content of an untrimmed sequence. These video-level labels are generally referred to as weak labels, and can be used to train AR models that identify and localize activities in untrimmed sequences.

In this paper, we propose a novel problem setting, named skeleton-based weakly-supervised temporal action localization (S-WTAL), which aims to detect actions in the skeleton stream and output beginning and end timestamps of the actions, as shown in Fig. 1. To the best of our knowledge, our study is the first to tackle the problem of S-WTAL.

Rather than extracting skeletons from videos after processing them via image-based methods, we propose to directly operate on skeleton motion clips. This is mainly due to the fact that in some cases, image-based videos may not necessarily focus on the human posture during actions. For example, in image-based AR for trimmed videos (Simonyan and Zisserman 2014; Feichtenhofer et al. 2019), some actions may be classified by the background without directly observing human action (He et al. 2016). Moreover, in the methods for untrimmed image-based videos (Singh and Lee 2017; Wang et al. 2017; Huang, Wang, and Li 2021), the untrimmed videos consist of several trimmed videos, which cause the depicted scenes to suddenly change from one to another, leading to disjoint fragments of human actions and providing a hint for the classification of different actions. Consequently, considering the subsequent applications such as animating various characters, the proposed S-WTAL setting is more suitable than image-based action localization.

However, compared with the image-based AR task, S-WTAL is more challenging. The difficulty stems from the fact that the model needs to learn body joint relations and activity dynamics solely from the skeleton data, where visual information of interaction with surrounding objects is not available. To overcome these challenges, we propose a framework that uses multiple instance learning (MIL) with frame-level label refinement (FLR) to learn the representations of the skeleton actions from video-level labels and refine the pseudo frame-level labels during model training.

To extract the spatio-temporal representations of each frame from the skeleton stream, we use spatio-temporal graph convolutional network (ST-GCN) (Yan, Xiong, and Lin 2018; Shi et al. 2019), which consists of graph convolutional networks (Kipf and Welling 2017) and temporal convolutional networks (Soo Kim and Reiter 2017). We train the ST-GCN with video-level labels by MIL, and then extract frame-level predictions as the pseudo labels of the frames. Since these frames are originally unlabeled, these pseudo labels would result in incorrect label assignments. Inspired by a solution to the noisy label problem (Tanaka et al. 2018), we update the network parameters and the pseudo labels alternately as a joint optimization to clean the noisy frame-level label predictions. At the same time, we adopt a two-headed network architecture with two independent classifiers to further reduce noisy samples by multi-view learning, and improve the quality of pseudo frame-level labels.

We evaluate our method on a skeleton-based AR benchmark, BABEL (Punnakkal et al. 2021). In many settings, our method outperforms existing methods by a large margin. We summarize the contributions of this paper as follows:

- We propose a novel problem setting S-WTAL, and a training methodology for the task.
- We propose a frame-level label refinement framework that extracts the action representation of each frame and refines the frame-level pseudo label by alternating optimization and multi-view learning.
- We evaluate our method by comparison with existing methods across several S-WTAL tasks, in which the proposal outperforms state-of-the-art by a considerable margin. Our approach successfully recognizes and localizes action segments in the skeleton streams.

Task	Input Type	Sequence Type	Label Type
AR	Image	Trimmed	Full
S-AR	Skeleton	Trimmed	Full
TAL	Image	Untrimmed	Full
WTAL	Image	Untrimmed	Weak
S-WTAL	Skeleton	Untrimmed	Weak

Table 1: Summary of action recognition tasks. Our proposed method is the only method applicable to untrimmed sequences of skeleton data with weak video-level labels.

Related Work

Skeleton-based Action Recognition

For skeleton-based AR, traditional methods use handcrafted features to model the human body (Vemulapalli, Arrate, and Chellappa 2014). After the breakthroughs in deep learning, data-driven methods consisting of Recurrent and Convolutional Neural Networks (RNNs and CNNs) have been widely used. RNN-based methods aim to model skeleton data as a sequence of coordinate vectors representing a human body joint (Du, Wang, and Wang 2015; Liu et al. 2016). CNN-based methods attempt to transform the skeleton data to pseudo images according to handcrafted transformation rules (Ke et al. 2017; Liu, Liu, and Chen 2017). However, these representations, vector sequences and 2D grids used by RNNs and CNNs, do not match the structure of skeleton data because they are naturally embedded as graphs. To address this issue, Yan *et al.*(Yan, Xiong, and Lin 2018) propose to directly model the skeleton data as graph structure with ST-GCN, which does not need handcrafted transformation rules. (Miki, Chen, and Demachi 2020) introduce local importance to ST-GCN in trimmed short sequences. Adaptive graph convolutional network (AGCN) (Shi et al. 2019) improves ST-GCN by parameterizing the graph structure of skeleton data and embedding it into the network.

However, all existing skeleton-based AR models are applied on trimmed sequences, which limits their utility in applications that require detailed scrutiny of the pose sequence, including action retrieval, intelligent surveillance, and human-computer interaction. Our method attempts to overcome this constraint to make skeleton-based AR applicable to untrimmed inputs.

Image-based Action Recognition

Among the tasks of video and action understanding, action classification of trimmed videos (Simonyan and Zisserman

2014; Feichtenhofer et al. 2019) is a popular task, generally named AR. Moreover, temporal action localization (TAL) (Chao et al. 2018; Shou, Wang, and Chang 2016) is another popular and challenging task. Conventional TAL methods require frame-level annotations (Chao et al. 2018; Lin et al. 2020). To avoid the labor-intensive and time-consuming task of manually assigning precise temporal annotation of each action instance, weakly-supervised TAL (WTAL) methods, which enable the network to be trained with video-label annotations, have drawn increasing attention (Singh and Lee 2017; Wang et al. 2017; Huang, Wang, and Li 2021). Hide-and-seek (Singh and Lee 2017) uses class activation maps to identify the relevant frames of each action. UntrimmedNet (Wang et al. 2017) introduces the MIL (Maron and Lozano-Pérez 1997; Paul, Roy, and Roy-Chowdhury 2018) framework to select foreground snippets and group them as action segments. STPN (Nguyen et al. 2018) improves upon UntrimmedNet by using a sparsity loss to encourage the sparsity of selected snippets. CoLA (Zhang et al. 2021) uses contrastive learning to identify foreground and background snippets. FAC-Net (Huang, Wang, and Li 2021) introduces a hybrid attention mechanism with class-wise foreground classification to improve the action localization performance. However, we note that all the methods introduced above are proposed to predict frame-level labels of image-based video streams.

Although the studies on skeleton-based and image-based AR have accomplished great achievements, there is no research trying to apply WTAL to skeleton data. Our method is the first trial for tackling S-WTAL, which aims at training a model to not only localize the action segments in untrimmed skeleton streams, but also achieve high localization performance when only the video-level annotations are available. With the development of S-WTAL, we can obtain clean action segments at low annotation cost, which benefits applications such as motion generation (Guo et al. 2020; Petrovich, Black, and Varol 2021). Table 1 summarizes the relationship between the proposed method and other existing approaches.

Method

In this section, we present our proposed framework for S-WTAL, which is shown in Fig. 2.

Problem Statement

S-WTAL aims to recognize and localize action segments in untrimmed 3D-skeleton videos given only video-level action labels during training. We assume that a video-label pair, $\{\boldsymbol{v}, \boldsymbol{y}\}$, drawn from a set of labeled videos with N samples $\{V, Y\} = \{(\boldsymbol{v}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$, is available. $\boldsymbol{v}$ denotes an untrimmed training video and $\boldsymbol{y} \in \mathbb{R}^C$ denotes its ground-truth label, where C is the number of action categories. Note that $\boldsymbol{y}$ could be a multi-hot vector if more than one action is present in the video, which means the action classification of a video is a multi-label classification. Since the input video is 3D-skeleton data, it is denoted by $\boldsymbol{v} \in \mathbb{R}^{T \times J \times 3}$, where T represents the length of the video, $J \times 3$ represents the position of the 25-joint ($J = 25$) skeleton used in NTU RGB+D (Shahroudy et al. 2016) in Cartesian coordinates, (x, y, z). We also denote the input as a movement sequence $\boldsymbol{v} = (\boldsymbol{v}^1, \cdots, \boldsymbol{v}^t, \cdots, \boldsymbol{v}^T)$. As the setting of S-WTAL, we only know the label set of $\boldsymbol{v}$, and do not know the frame-level label of $\boldsymbol{v}^t$. The goal of our method is to correctly predict the frame-level label of each $\boldsymbol{v}^t$ and then extract action segments by thresholding the frame-level predictions, as described by the implementational details in the Experiment Section.

Feature Extraction

To extract spatio-temporal features from the skeleton data, we use the modules before the global average pooling (GAP) layer of AGCN proposed in (Shi et al. 2019) as the feature extractor E. We add a temporal convolutional block with stride 2 between the first and the second block of AGCN to obtain the temporal representations. Differing from WTAL methods based on image data that divide each untrimmed video into non-overlapping 16-frame snippets, we directly input the whole video $\boldsymbol{v}$ with $T \times J \times 3$ dimensions into the feature extractor and obtain the representations $\boldsymbol{f} = E(\boldsymbol{v}) \in \mathbb{R}^{\frac{T}{8} \times D}$, where $D = 256$ is the feature dimensionality. Since the goal of S-WTAL is to generate the action prediction of each frame, we further up-sample the feature $\boldsymbol{f}$ into $\mathbb{R}^{T \times D}$ by linear interpolation along the temporal dimension and apply L2 normalization along the feature dimension.

Action Localization

To achieve S-WTAL, we train the classifier with only video-level annotations by MIL. We build an extended action classifier as a cosine classifier consisting of weight vectors $\mathbf{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_C, \boldsymbol{w}_{C+1}]$, where the first C classes correspond to the action categories, and the $C + 1$ class the background.

When the feature $\boldsymbol{f}$ is obtained, cosine similarities between $\boldsymbol{f}$ and the weights in $\mathbf{W}$ are calculated to obtain the frame-level class activation scores $\boldsymbol{s} \in \mathbb{R}^{T \times (C+1)}$:

$$\boldsymbol{s}^{t,c} = \delta \cdot \cos(\boldsymbol{f}^t, \boldsymbol{w}_c), \tag{1}$$

where $\boldsymbol{f}^t$, $\boldsymbol{w}_c$ and δ represent the feature of frame t, the weight of class c, and a scalar to control the scale of the value, respectively.

To perform MIL, we calculate the class-wise attention scores $\boldsymbol{a} \in \mathbb{R}^{T \times (C+1)}$ by the softmax function along the temporal dimension, which can be written as:

$$\boldsymbol{a}^{t,c} = \frac{\exp(\tau \cdot \boldsymbol{s}^{t,c})}{\sum_{k=1}^{T} \exp(\tau \cdot \boldsymbol{s}^{k,c})}, \tag{2}$$

where τ denotes a temperature parameter to control the smoothness of the softmax function. Then, the video-level class activation scores $\boldsymbol{r} \in \mathbb{R}^{C+1}$ can be calculated by aggregating the frame-level class activation scores $\boldsymbol{s}$ according to the attention $\boldsymbol{a}$ as follows:

$$\boldsymbol{r}^c = \sum_{t=1}^{T} \boldsymbol{a}^{t,c} \boldsymbol{s}^{t,c}. \tag{3}$$

Since the video-level classification in S-WTAL is a multi-label problem, we apply sigmoid function for each class:

$$\hat{\boldsymbol{r}}^c = \mathrm{Sigmoid}(\boldsymbol{r}^c), \tag{4}$$

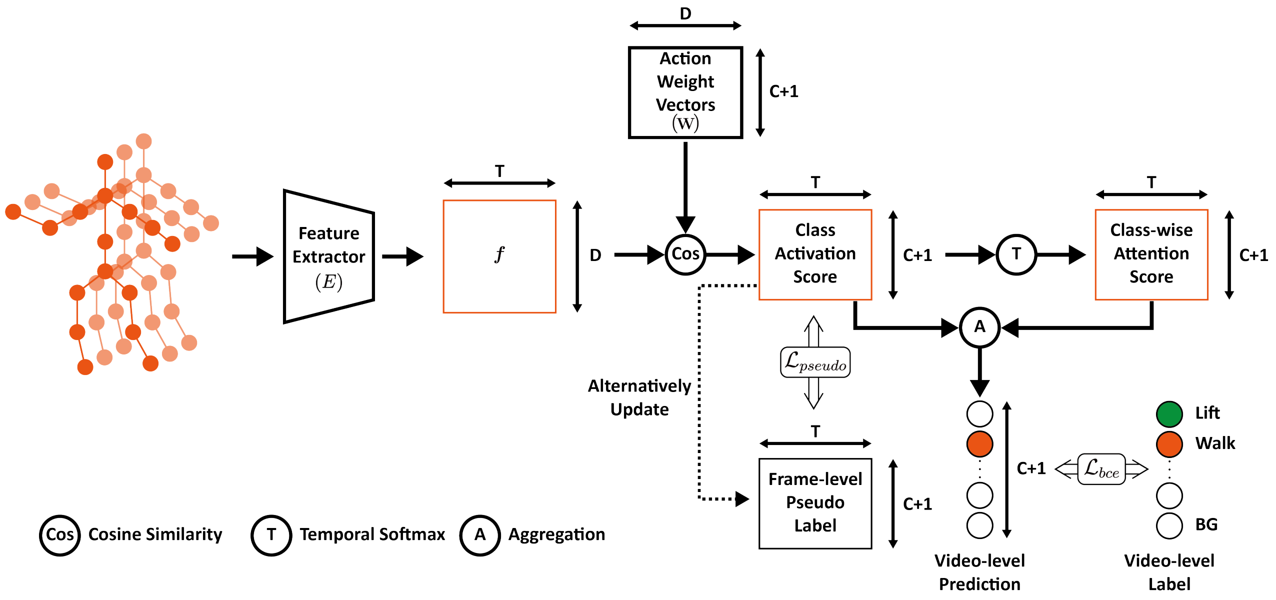

Figure 2: Overview of the proposed framework, which has one feature extractor (E) and an action classifier with weight vectors ($\mathbf{W}$). The network is trained via the MIL under the supervision of video-level labels and the FLR of frame-level pseudo labels.

Algorithm 1: Joint Optimization

```
for $e \leftarrow 0$ to $\#Epoch$ do
    update $\boldsymbol{\theta}^{(e+1)}$ by Adam on $\mathcal{L}_{bce} + \mathcal{L}_{pseudo} + \mathcal{L}_{KL}$
    update $\bar{\boldsymbol{s}}^{(e+1)}$ as $\bar{\boldsymbol{s}}^{(e+1)} \leftarrow \hat{\boldsymbol{s}}^{(e)}$ by Eq. (9)
end for
```

and then train the model with binary cross-entropy loss:

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C+1} \boldsymbol{y}_i^c \log \hat{\boldsymbol{r}}_i^c + (1-\boldsymbol{y}_i^c)\log(1-\hat{\boldsymbol{r}}_i^c), \tag{5}$$

where $\boldsymbol{y}_i$ denotes the ground truth label of the i-th sample. Because the background frames exist in all videos, all videos should have the background label, *i.e.*, $\boldsymbol{y}_i^{C+1} = 1$.

Frame-Level Label Refinement

Joint Optimization Through training with MIL, the model can localize the action segments to some extent. However, we found that the performance of localization increases at the beginning of the training, and instead of showing further improvement, eventually decreases during the training process. Since only video-level annotations are available for training the model, the trained model tends to only recognize the video-level actions according to the key frames of each action. This results in detected action segments to contain only the salient key frames and not the related frames that constitute each action, which inevitably leads to a decrease in the performance of action localization.

To solve this problem, we interpret the generation of frame-level labels in S-WTAL as a noisy label problem. When we train the model with MIL, we can obtain the frame-level predictions by applying softmax function to frame-level class activation scores as follows:

$$\hat{\boldsymbol{s}}^{t,c} = \frac{\exp(\boldsymbol{s}^{t,c})}{\sum_{k=1}^{C+1}\exp(\boldsymbol{s}^{t,k})}. \tag{6}$$

As previously mentioned, we observed that these predictions are correct to some extent at the beginning of the training, but start to overfit to key frames during the training process. This phenomenon is similar to that observed in the noisy label problem (Tanaka et al. 2018).

Suppose the pseudo labels for each frame are available in some way, we can train the model with the frame-level pseudo labels with the cross-entropy loss as follows:

$$\mathcal{L}_{pseudo} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C+1} \bar{\boldsymbol{s}}_i^{t,c}\log\hat{\boldsymbol{s}}_i^{t,c}, \tag{7}$$

where $\bar{\boldsymbol{s}}_i^{t,c}$ denotes the pseudo label for the c-th class of the t-th frame in the i-th video and may contain some noise.

To deal with the noisy label problem, Tanaka et al. (Tanaka et al. 2018) observed that if a network is trained with a high learning rate, it is less likely to overfit to noisy labels. Consequently, the loss Eq. (7) is high for noisy labels and low for clean labels, which suggests that clean pseudo labels can be obtained by updating the pseudo labels in the direction that decreases Eq. (7). To achieve this, we formulate the problem as the joint optimization of the network parameters and the pseudo labels:

$$\min_{\boldsymbol{\theta},\bar{\boldsymbol{s}}} \mathcal{L}_{pseudo}, \tag{8}$$

where $\boldsymbol{\theta}$ denotes the network parameters of the feature extractor E and the action weight vectors $\mathbf{W}$ in the classifier. Alternately updating the network parameters and pseudo labels is achieved via joint optimization (Tanaka et al. 2018) by repeating these two steps:

Updating $\boldsymbol{\theta}$ with fixed $\bar{\boldsymbol{s}}$: As all terms in the loss function Eq. (7) are differentiable with respect to $\boldsymbol{\theta}$, we update $\boldsymbol{\theta}$ by the Adam optimizer (Kingma and Ba 2014) on Eq. (7).

Updating $\bar{\boldsymbol{s}}$ with fixed $\boldsymbol{\theta}$: Considering the update of $\bar{\boldsymbol{s}}$, we need to minimize $\mathcal{L}_{pseudo}$ with fixed $\boldsymbol{\theta}$ to correct the pseudo labels. $\mathcal{L}_{pseudo}$ can be minimized when the predictions of the network equals $\bar{\boldsymbol{s}}$. As a result, $\bar{\boldsymbol{s}}$ is updated as follows:

$$\bar{\boldsymbol{s}}^{(e+1)} \leftarrow \hat{\boldsymbol{s}}^{(e)}, \tag{9}$$

which means the frame-wise predictions of the current epoch e are used as the pseudo labels for the next epoch $e+1$ to reduce the noise of the pseudo label before the network overfits to noisy labels. The entire algorithm of this joint optimization is as described in Algorithm 1.

Multi-View Learning To improve the robustness of the model to noisy labels, we also apply the multi-view learning strategy inspired by co-training (Blum and Mitchell 1998; Zhang et al. 2018; Han et al. 2018) to the model training process. Specifically, instead of using a single action classifier, we prepare two action classifiers with different weight vectors $\mathbf{W}_1$ and $\mathbf{W}_2$. We initialize the two classifiers with different random initial parameters, but train them with the same data at the mini-batch level. As a result, the two independently trained classifiers have different abilities to learn from the noisy labels, which means the two classifiers tend to output similar results from clean samples and dissimilar results from noisy ones (Wei et al. 2020; Yu, Hashimoto, and Ushiku 2021). To eliminate the effects of noisy samples, we calculate the agreement of the two classifiers by the symmetric Kullback Leibler (KL) divergence between predictions of the two networks $\hat{\boldsymbol{s}}_1$ and $\hat{\boldsymbol{s}}_2$:

$$l_{KL} = D_{\mathrm{KL}}(\hat{\boldsymbol{s}}_1^t || \hat{\boldsymbol{s}}_2^t) + D_{\mathrm{KL}}(\hat{\boldsymbol{s}}_2^t || \hat{\boldsymbol{s}}_1^t), \tag{10}$$

where

$$D_{\mathrm{KL}}(\hat{\boldsymbol{s}}_1 || \hat{\boldsymbol{s}}_2) = \sum_{c=1}^{C+1} \hat{\boldsymbol{s}}_1^{t,c} \log \frac{\hat{\boldsymbol{s}}_1^{t,c}}{\hat{\boldsymbol{s}}_2^{t,c}}, \tag{11}$$

$$D_{\mathrm{KL}}(\hat{\boldsymbol{s}}_2 || \hat{\boldsymbol{s}}_1) = \sum_{c=1}^{C+1} \hat{\boldsymbol{s}}_2^{t,c} \log \frac{\hat{\boldsymbol{s}}_2^{t,c}}{\hat{\boldsymbol{s}}_1^{t,c}}. \tag{12}$$

To learn from clean samples, we select examples having small l_{KL}, where the selected small-loss instances are more likely to be with clean labels. Specifically, we conduct small-loss selection:

$$S' = \arg\min\nolimits_{S':|S'| \geq \alpha|S|} l_{KL}(S), \tag{13}$$

where S denotes the frame-level predictions of the dataset by $S = \{\{\hat{\boldsymbol{s}}_i^t\}_{t=1}^{T}\}_{i=1}^{N}$. This equation indicates that we only use $\alpha\%$ frames in the dataset to minimize the divergence. The average loss on these examples are calculated as:

$$\mathcal{L}_{KL} = \frac{1}{|S'|} \sum_{\boldsymbol{s} \in S'} l_{KL}(\boldsymbol{s}). \tag{14}$$

Overall Objective Function

In summary, our frame-level label refinement framework performs MIL, joint optimization, and multi-view learning. The overall learning objective is:

$$\min_{\boldsymbol{\theta}, \bar{\boldsymbol{s}}} \mathcal{L}_{bce} + \lambda_{pseudo}\mathcal{L}_{pseudo} + \lambda_{KL}\mathcal{L}_{KL}. \tag{15}$$

Dataset	Actions	#Training Sequences	#Test Sequences
Subset-1	walk, stand, turn, jump	2,954	1,001
Subset-2	sit, run, stand-up, kick	624	212
Subset-3	jog, wave, dance, gesture	232	102

Table 2: Details of the subsets used in the experiment.

Experiment

Experimental Setup

Datasets. We verify the effectiveness of our approach on the benchmark dataset of 3D human motion, BABEL (Punnakkal et al. 2021). BABEL is the only dataset containing large-scale 3D human motion sequences with frame-level labels, which describe all actions in every frame of the sequences. BABEL annotates about 43 hours of mocap sequences from AMASS (Mahmood et al. 2019), which consist of over 63k frame-level labels and over 250 unique action categories. We generate the video-level labels by gathering the frame-level labels of each sequence. Due to the difficulty of S-WTAL, we create 3 subsets, each of which consists of 4 action categories. The details of each subset are shown in Table 2.

Comparison of Methods. Since there is no existing method for solving S-WTAL directly, we combine existing WTAL methods with the skeleton-based feature extractor for comparison. As the proposed frame-level label refinement is built with AGCN (Shi et al. 2019) as the feature extractor, we also apply AGCN to WTAL methods. Specifically, we incorporate two state-of-the-art WTAL methods CoLA (Zhang et al. 2021) and FAC-Net (Huang, Wang, and Li 2021) based on AGCN to provide fair and meaningful comparison.

Evaluation Protocols. We adopt the same standard metrics for evaluating the performance of AR and localization as WTAL *i.e.*, mean Average Precisions (mAPs) under different Intersection of Union (IoU) thresholds, by transforming the frame-level predictions to action segments. In practice, we adopt the official evaluation code provided by FAC-Net (Huang, Wang, and Li 2021). Additionally, we also evaluate the classification mAP at the video level to verify the effectiveness of our approach.

Implementation Details. We use the PyTorch library (Paszke et al. 2019) to implement the proposed framework on a single NVIDIA A100 GPU. The entire network, including one feature extractor and two classifiers, is jointly trained in an end-to-end manner. Except for the modification mentioned in the section of feature extraction, the feature extractor uses the same parameter setting as in the original papers. The Adam optimizer with the learning rate of 0.0001 is applied to optimize the network at the mini-batch level with batch-size 8 for 100 epochs. The hyper-parameters adopted to construct the classifier are empirically set as follows: $\delta = 5.0, \tau = 2.0$. The α used for selecting clean samples in Eq. (13) is set as 0.2. The loss weights λ_{pseudo} and λ_{KL} are set to 1.0 and 0.5, respectively. We begin updating the frame-level pseudo labels from the 10th epoch, and we use the average output probability for each frame from the past

Subset-1

Method	Classification	Detection mAP @ IoU (%)					
	mAP (%)	0.1	0.2	0.3	0.4	0.5	Avg
CoLA	73.84 ± 1.03	27.40	14.63	6.43	4.03	2.15	10.93 ± 0.53
FAC-Net	76.74 ± 1.53	29.65	17.65	8.48	3.88	2.62	12.46 ± 0.63
Ours	**80.21 ± 0.68**	**48.74**	**39.82**	**33.15**	**27.39**	**21.70**	**34.16 ± 1.26**
Ours w/o FLR	77.06 ± 0.43	25.92	15.02	7.81	4.48	2.17	11.08 ± 1.08

Subset-2

Method	Classification	Detection mAP @ IoU (%)					
	mAP (%)	0.1	0.2	0.3	0.4	0.5	Avg
CoLA	82.96 ± 1.50	40.10	26.86	19.90	14.02	10.32	22.24 ± 2.41
FAC-Net	86.47 ± 1.18	34.18	19.84	12.95	9.03	6.53	16.51 ± 2.75
Ours	**89.15 ± 1.59**	**61.01**	**50.26**	**40.36**	**29.84**	**19.55**	**40.20 ± 3.09**
Ours w/o FLR	83.97 ± 0.91	40.30	31.31	18.82	11.52	5.16	21.42 ± 2.89

Subset-3

Method	Classification	Detection mAP @ IoU (%)					
	mAP (%)	0.1	0.2	0.3	0.4	0.5	Avg
CoLA	34.74 ± 5.43	21.50	17.40	15.16	12.10	8.99	15.03 ± 1.78
FAC-Net	59.90 ± 6.12	27.51	22.26	17.64	14.05	8.90	18.07 ± 4.25
Ours	**67.08 ± 3.34**	**35.81**	**31.45**	**26.55**	**23.42**	**20.36**	**27.52 ± 1.23**
Ours w/o FLR	65.90 ± 2.86	19.52	12.65	7.80	3.87	1.61	9.09 ± 1.59

Table 3: Video-level classification and segment-level detection performance comparisons over BABEL. The column Avg indicates the average mAP at IoU thresholds from 0.1 to 0.5. The averages and standard deviations are obtained from three trials.

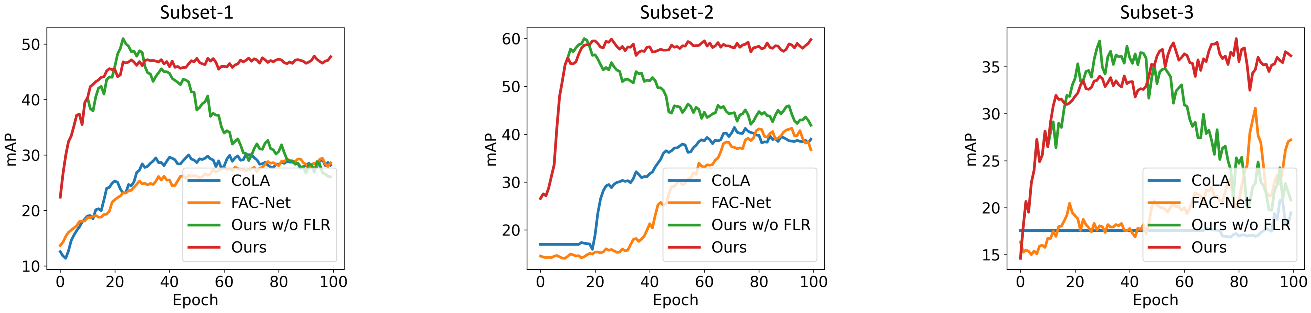

Figure 3: Detection mAP@0.1 vs. epoch curve of each method on each subset.

10 epochs as $\bar{s}$ in Eq. (9).

During inference, we reject the category whose class probability $\hat{\boldsymbol{r}}^c$ in Eq. (5) is lower than 0.1. Following (Huang, Wang, and Li 2021), a set of thresholds are used to obtain the predicted action segments, then non-maximum suppression is performed to remove overlapping segments.

Experimental Results

The classification mAP of each video and the detection mAP of each detected action segment on the three subsets are shown in Fig. 3. To evaluate the performance of video-level multi-label classification, we report the mAP by averaging the AP of each class. Regarding the action localization performance, since S-WTAL is still a challenging task, we report the mAP at thresholds from 0.1 to 0.5 and their average. According to Fig. 3, we find that our method outperforms the existing approaches in all the metrics by a large margin. We also report the proposed method without FLR, and its performance drops significantly, indicating the effectiveness of the proposed frame-level label refinement.

We show the learning curve of each method on each subset in Fig. 3. Despite using the features extracted from the same base network, the proposed framework far outperforms existing WTAL methods CoLA and FAC-Net. This demonstrates that the task of S-WTAL is much more challenging than that of image-based WTAL. As mentioned in the section of frame-level label refinement, it is interesting to observe that the proposed framework without FLR achieves good results at the beginning of the training, but then starts to overfit to key frames during the training process, lowering the performance. However, our method with FLR can maintain high performance until the end of the training process.

Some examples of detected action segments are visualized in Fig. 4. In the first example of sitting, our method

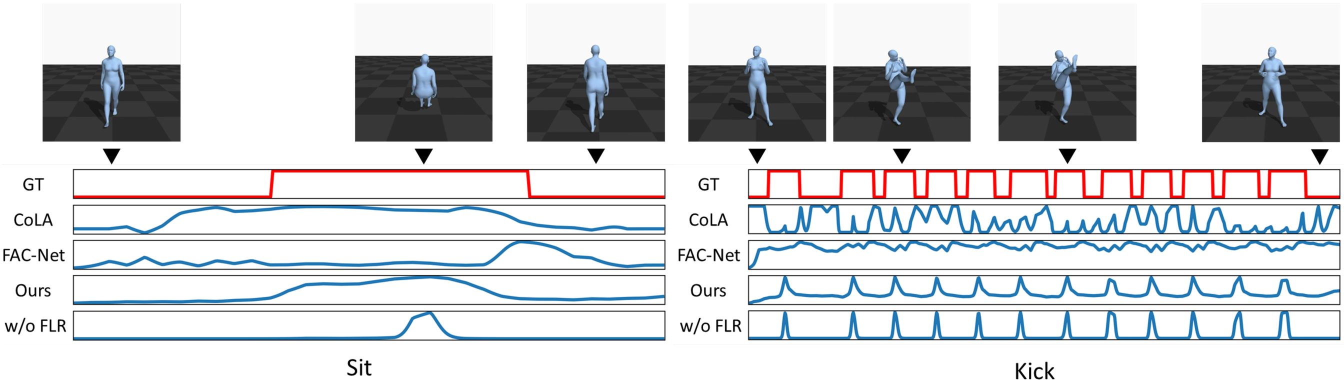

Figure 4: Qualitative results on BABEL. We show the ground truth and the activation scores of each method.

Dataset	Subset-1				Subset-2				Subset-3			
	Walk	Stand	Turn	Jump	Sit	Run	Stand-up	Kick	Jog	Wave	Dance	Gesture
CoLA	43.87	37.98	14.68	13.07	78.40	40.35	19.20	22.45	23.03	5.93	43.67	13.36
FAC-Net	49.17	31.65	22.52	15.24	57.11	36.59	16.44	26.57	40.45	15.28	42.56	11.76
Ours	79.34	**74.32**	24.28	17.03	83.74	**57.49**	**23.59**	**79.21**	45.45	16.93	**66.13**	14.75
w/o FLR	44.07	45.23	5.39	9.01	56.89	45.02	6.03	53.29	41.80	7.27	22.63	6.37
w/o JO	55.73	56.07	8.89	13.57	60.18	46.75	8.99	55.16	42.34	9.82	25.91	10.37
w/o MVL	78.14	67.09	16.67	**20.91**	81.09	56.49	20.42	79.04	43.92	15.39	62.37	13.65
$\lambda_{pseudo} = 0.2$	77.82	69.24	19.25	12.99	82.16	51.89	21.14	80.14	**48.48**	18.17	44.90	11.16
$\lambda_{pseudo} = 0.5$	78.72	68.55	18.87	14.63	83.32	55.97	20.66	80.03	44.93	16.89	65.71	17.59
$\lambda_{KL} = 0.2$	78.91	73.02	21.00	18.70	**84.10**	50.80	22.57	79.38	43.85	17.44	63.15	12.26
$\lambda_{KL} = 1.0$	79.27	72.92	23.21	16.11	83.05	54.05	20.34	79.39	44.75	**19.29**	60.66	18.55
$\alpha = 0.1$	79.36	73.97	19.97	16.81	81.60	54.31	22.64	79.77	43.34	17.02	59.86	18.00
$\alpha = 0.5$	**79.37**	73.35	**24.56**	15.93	82.58	52.17	22.62	**80.19**	47.63	16.57	52.42	**20.26**

Table 4: Results of analysis tasks. Detection mAP@0.1 are reported.

shows high action scores when the action occurs. In the second example of kicking, where the action is frequently repeated in the video, the proposed method successfully detects all action segments, indicating that our method can handle dense action occurrences. Furthermore, our proposal, FLR, prevents our model from overfitting to key frames, providing better coverage of the ground truth.

Analysis

Variants of the proposed method were evaluated using the BABEL dataset, for further exploration of the efficacy of the proposal. The following variants were studied: (1) "Ours w/o FLR" is a variant that does not use FLR, *i.e.*, $\mathcal{L}_{pseudo}$ and $\mathcal{L}_{KL}$ in Eq. (15). (2) "Ours w/o JO" is a variant that does not use joint optimization on frame-level pseudo labels, *i.e.*, $\mathcal{L}_{pseudo}$ in Eq. (15). (3) "Ours w/o MVL" is a variant that omits the multi-view learning loss $\mathcal{L}_{KL}$ in Eq. (15).

Table 4 shows the comparisons of the detection mAP on each class of each subset between the existing methods and the variants of our method. The results reveal that the version of our approach that uses all the losses outperforms other variants in all settings on average. Specifically, the most important component for our method is $\mathcal{L}_{\mathrm{pseudo}}$, and $\mathcal{L}_{\mathrm{KL}}$ is also necessary to achieve higher performance. In Table 4, we also demonstrate the sensitivity of the loss in Eq. (15) to different weights and the α. We find that our method is robust to these hyper-parameters.

Conclusion

In this study, we proposed a frame-level label refinement framework for a novel AR problem setting, S-WTAL. Our framework uses AGCN to learn the spatio-temporal features from the skeleton data, and generates the frame-level pseudo labels by MIL with video-level annotations. The frame-level pseudo labels are further refined by joint optimization and multi-view learning. The performance of the proposed method was evaluated on a real dataset across various settings, and our method outperformed existing state-of-the-art WTAL methods by a considerable margin. To improve the detection performance of some difficult motion classes, we would like to analyze the characteristics of each motion to better estimate the duration of each action segment in future work.

References

Blum, A.; and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *COLT*.

Chao, Y.-W.; Vijayanarasimhan, S.; Seybold, B.; Ross, D. A.; Deng, J.; and Sukthankar, R. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *ACMMM*.

Habermann, M.; Xu, W.; Zollhoefer, M.; Pons-Moll, G.; and Theobalt, C. 2019. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics*.

Habermann, M.; Xu, W.; Zollhofer, M.; Pons-Moll, G.; and Theobalt, C. 2020. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *NeurIPS*.

Hasler, N.; Rosenhahn, B.; Thormahlen, T.; Wand, M.; Gall, J.; and Seidel, H.-P. 2009. Markerless motion capture with unsynchronized moving cameras. In *CVPR*.

He, Y.; Shirakabe, S.; Satoh, Y.; and Kataoka, H. 2016. Human action recognition without human. In *ECCV-W*.

Huang, L.; Wang, L.; and Li, H. 2021. Foreground-Action Consistency Network for Weakly Supervised Temporal Action Localization. In *CVPR*.

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A new representation of skeleton sequences for 3d action recognition. In *CVPR*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Lin, C.; Li, J.; Wang, Y.; Tai, Y.; Luo, D.; Cui, Z.; Wang, C.; Li, J.; Huang, F.; and Ji, R. 2020. Fast learning of temporal action proposal via dense boundary generator. In *AAAI*.

Liu, J.; Shahroudy, A.; Xu, D.; and Wang, G. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*.

Liu, M.; Liu, H.; and Chen, C. 2017. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *CVPR*.

Maron, O.; and Lozano-Pérez, T. 1997. A framework for multiple-instance learning. In *NeurIPS*.

Miki, D.; Chen, S.; and Demachi, K. 2020. Weakly supervised graph convolutional neural network for human action localization. In *WACV*.

Nguyen, P.; Liu, T.; Prasad, G.; and Han, B. 2018. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Paul, S.; Roy, S.; and Roy-Chowdhury, A. K. 2018. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3D Human Motion Synthesis with Transformer VAE. In *ICCV*.

Punnakkal, A. R.; Chandrasekaran, A.; Athanasiou, N.; Quiros-Ramirez, A.; and Black, M. J. 2021. BABEL: Bodies, action and behavior with english labels. In *CVPR*.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*.

Shou, Z.; Wang, D.; and Chang, S.-F. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*.

Singh, K. K.; and Lee, Y. J. 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*.

Soo Kim, T.; and Reiter, A. 2017. Interpretable 3d human action analysis with temporal convolutional networks. In *CVPR-W*.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *CVPR*.

Tang, Y.; Tian, Y.; Lu, J.; Li, P.; and Zhou, J. 2018. Deep progressive reinforcement learning for skeleton-based action recognition. In *CVPR*.

Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*.

Wang, L.; Xiong, Y.; Lin, D.; and Van Gool, L. 2017. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.

Yu, Q.; Hashimoto, A.; and Ushiku, Y. 2021. Divergence Optimization for Noisy Universal Domain Adaptation. In *CVPR*.

Zhang, C.; Cao, M.; Yang, D.; Chen, J.; and Zou, Y. 2021. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *CVPR*.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *CVPR*.

Zheng, N.; Wen, J.; Liu, R.; Long, L.; Dai, J.; and Gong, Z. 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *AAAI*.