Prompt-enhanced hierarchical transformer elevating cardiopulmonary resuscitation instruction via temporal action segmentation

Yang Liu [a], Xiaoyun Zhong [a,b], Shiyao Zhai [a,b], Zhicheng Du [a,b], Zhenyuan Gao [c], Qiming Huang [d], Can Yang Zhang [a], Bin Jiang [e], Vijay Kumar Pandey [a,b], Sanyang Han [a], Runming Wang [a], Yuxing Han [f], Chuhui Wang [a,b], Peiwu Qin [a,b,*]

[a] *Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen, 518055, China*
[b] *Center of Precision Medicine and Healthcare, Tsinghua-Berkeley Shenzhen Institute, Shenzhen, 518055, China*
[c] *Miragestars Inc., Tianjin, 300392, China*
[d] *Shenzhen ZNV Technology Co., Ltd, Shenzhen, 518057, China*
[e] *School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, 264209, China*
[f] *Tsinghua Shenzhen International Graduate School, Shenzhen, 518055, China*

ARTICLE INFO

ABSTRACT

The vast majority of people who suffer unexpected cardiac arrest are performed cardiopulmonary resuscitation (CPR) by passersby in a desperate attempt to restore life, but endeavors turn out to be fruitless on account of disqualification. Fortunately, many pieces of research manifest that disciplined training will help to elevate the success rate of resuscitation, which constantly desires a seamless combination of novel techniques to yield further advancement. To this end, we collect a specialized CPR video dataset in which trainees make efforts to behave resuscitation on mannequins independently in adherence to approved guidelines, promoting an auxiliary toolbox to assist supervision and rectification of intermediate potential issues via modern deep learning methodologies. Our research empirically views this problem as a temporal action segmentation (TAS) task in computer vision, which aims to segment an untrimmed video at a frame-wise level. Here, we propose a Prompt-enhanced hierarchical Transformer (PhiTrans) that integrates three indispensable modules, including a textual prompt-based Video Features Extractor (VFE), a transformer-based Action Segmentation Executor (ASE), and a regression-based Prediction Refinement Calibrator (PRC). The backbone preferentially derives from applications in three approved public datasets (GTEA, 50Salads, and Breakfast) collected for TAS tasks, which experimentally facilitates the model excavation on the CPR dataset. In general, we probe into a feasible pipeline that elevates the CPR instruction qualification via action segmentation equipped with novel deep learning techniques. Associated experiments on the CPR dataset advocate our resolution with surpassing 91.0% on Accuracy, Edit score, and F1 score.

1. Introduction

Out-of-hospital cardiac arrest (OHCA) is a universal public health issue undergone by about 3.8 million people annually, with only 8% to 12% surviving hospital discharge [1]. Characterized as blood flow or breathing stops, OHCA induces permanent brain damage or death happens acutely. Performing cardiopulmonary resuscitation (CPR) could serve as an emergency procedure for OHCA, which maintains the blood flow and breathing until advanced medical help arrives [2]. There has been a large volume of research and practice for decades to investigate CPR [3–5]. In 1891, the first chest compression on a human being were performed by Friedrich Maass [4]. The first guidelines for CPR were released about 50 years ago [3]. Despite a long history of

deploying CPR against OHCA, survival remains dismally low. There are indications that CPR performance influences the outcome [6,7]. From this perspective, a great deal of research [8–10] seeks and proves the positive effects of CPR instruction. Besides, the increasing number of people suffering OHCA worldwide makes intensive CPR education even more imperative. To enable CPR education as a mandatory part of society, not only should we cultivate the awareness of social responsibility, but employ more comprehensive approaches [8]. Specifically, traditional assessment of CPR skills involves strenuous manual efforts, which lacks efficiency and repeatability [8,9]. Extensive attempts to revive those who sustain OHCA will be probably further improved

---

with the combination of promising novel computer techniques and widespread application.

One most relevant specific research [11] emerges by collecting a dataset of real-world instruction videos from the Internet, containing performing CPR and four non-medical tasks. However, the work intends to address the problem of automatically learning the main steps to complete a certain task, neither implemented particularly for formal CPR behaviors nor designed for CPR instruction. Given the aforementioned dataset, several subsequent works focus on action segmentation in an unsupervised [12] or weakly supervised way [13] without proposing concrete challenges pertaining to CPR instruction. The positive impacts of CPR action segmentation are not clearly defined yet. Therefore, we evoke one ensuing challenge: *How better could we elevate CPR instruction with action segmentation?*

To this end, we resort to investigating a specific realm called temporal action segmentation (TAS), which has gradually developed into one of the high-profile research spotlights in computer vision. Extensive researches [14–17] on medical image segmentation have been conducted in recent years. Compared with these image segmentation tasks, the universal goal of the TAS task is to identify activities in untrimmed videos at a frame-wise level. It has promoted a wealth of applications in human behavior analysis from video summarization [18], video surveillance [19], action recognition [20,21], to skill assessment [22]. With prosperous computing power reinforcement, understanding single-semantics short video clips has been gradually outmoded in the TAS task in favor of larger, more complex untrimmed videos [23,24], which requires both intrinsic and extrinsic correlations of actions. Conventional segmentation methods [25] like Temporal Convolutional Networks (TCNs) consider single frames or short video segments for feature representation. They overlook the latent relationship among contextual actions, leading to poor performance, especially in long videos. Accordingly, some studies [26,27] exploit Recurrent Neural Networks (RNNs) to model each action clip to maintain local dependencies but still struggle to handle longer videos effectively due to the inherent spatio-temporal complexity of RNN. To equip the model with relational reasoning, methods utilizing Graph Convolutional Networks (GCNs) regard each action as a single node on the graph and edges represent the contextual relationship [28,29]. However, these preceding works all adopt frame-wise features extracted by pre-trained I3D [30] network, which might not be adequate enough to excavate effective representations of videos. In the past few years, it can be witnessed that both transformer-based [31,32] and prompt-based architectures [33,34] have flourished in artificial intelligence, which tremendously lightens our research on addressing deficiencies of preceding methods. Particularly in visual applications, Transformer-based architectures hold the potential to integrate the information between sequential elements that are far from each other with powerful scalability. Prompt-based architectures could enhance the visual features with representative linguistic semantics.

Now we attempt to yield a feasible resolution for the proposed challenge: Firstly, a specialized dataset dedicated solely to CPR actions is crucial. However, existing action segmentation datasets only contain limited videos with CPR actions. Moreover, the quality of these videos varies, and the associated annotations are not comprehensive nor professional. To bridge this gap, we preliminarily establish a specialized CPR dataset involving videos of participants performing the whole process of CPR in a standard green screen laboratory environment. These videos contain 15 CPR categories and are labeled at a frame level by experienced trainers. After that, we devise a crafted architecture to perform action segmentation. Specifically, we propose a Prompt-enhanced hierarchical Transformer (PhiTrans) that integrates three integral modules: (1) a textual prompt-based Video Features Extractor (VFE) module that extracts abundant frame-wise features; (2) a transformer-based Action Segmentation Executor (ASE) module that deduces the contextual relationship while adaptive to long frame sequences; (3) a regression-based Prediction Refinement Calibrator

(PRC) module that further alleviates the over-segmentation issues highlighted in the TAS task. Finally, we observe the model performance on the custom CPR dataset and claim that it can serve as an auxiliary toolbox applied in action segmentation for assisting CPR instruction by automatically identifying potential omission, repetition, or out-of-order situations at a frame-wise level, allowing trainees to rectify the workflow free of experts.

Our main contributions are chronologically three-fold:

- We reveal and fill the gap of lacking a specialized CPR dataset adequate for action segmentation. The collected CPR dataset contains videos of the entire cardiopulmonary resuscitation evaluation and associated frame-level semantic annotations.
- We propose an integrated model, called PhiTrans, especially applied for CPR action segmentation, including three integral modules: Video Features Extractor, Action Segmentation Executor, and Prediction Refinement Calibrator.
- To the best of our knowledge, we unprecedentedly probe into a feasible pipeline that genuinely elevates the CPR instruction qualification via action segmentation in conjunction with cutting-edge deep learning techniques.

2. Related work

2.1. Temporal action segmentation

Temporal action segmentation (TAS), one of the most challenging topics in advanced video comprehension, aims to extract frame-wise features from untrimmed videos and categorize them chronologically with pre-defined action labels. Action localization, video summarization, and other downstream applications benefit from products of the TAS task as input. Over the past decade, a multitude of methods [25,35–38] have leveraged models for action prediction with the extracted frame-wise features. Traditional paradigms involve sliding windows with non-maximum suppression [35], Hidden Markov Models (HMM) [36], Linear Dynamical Systems (LDS) [37], and Bayesian Non-parametric Models (BNM) [25].

These methods encounter common obstacles to modeling long-range dependencies. To alleviate this issue, RNNs [27] and TCNs [38] are deployed to capture global dependencies and the contextual information of adjacent frames. After that, Multi-Stage Temporal Convolutional Network (MS-TCN) [24] combines TCNs with multi-stage patterns to make remarkable progress in the TAS task, wherein plural stages are stacked to refine the predictions from the preceding output successively. In addition, other methods aim to model the TAS task from a unique perspective. Self-Supervised Temporal Domain Adaptation (SSTDA) [39] trains with two auxiliary tasks of binary and sequential domain prediction. Dilated Temporal Graph Reasoning Module (DTGRM) [28] builds multi-level dilated temporal graphs to simulate temporal dependencies between video frames at different timescales. Unified Video Action Segmentation model via Transformers (UVAST) [40] presents a unified design for fully and timestamp supervised temporal action segmentation via Transformers in a seq2seq style. Action Shared-Private network (ASPnet) [41] fuses multiple data sources and captures long-range temporal dynamics in sequential data. These especial methods lose portability on account of complex task patterns. Considering the spatio-temporal complexity of CPR action segmentation and accessibility for downstream design, we follow the same philosophy of multi-stages in previous methods [24,42] for iterative refinement.

2.2. Transformer-based architecture

Initially designed for Natural Language Processing (NLP) related tasks, Transformer [31] has motivated a tremendous leap forward in capabilities for pre-training on larger datasets and fine-tuning on smaller task-specific datasets with computational efficiency and scalability. The

visual application [32] of Transformer constantly challenges the dominant status of Convolutional Neural Networks (CNNs). Vision Transformers require less vision-specific inductive bias and maintain more global information relying on the Multi-head Self-Attention (MSA).

Witnessed the success of Transformer in image classification, image segmentation, and other vision tasks, one study [43] explores the transfer implementation of Transformer on the TAS task. The work presents major concerns underwent and accordingly proposes local connectivity inductive bias and hierarchical representation pattern, allowing vanilla Transformer scale to the TAS task. Such adaptions produce a hybrid Transformer architecture with temporal convolution included, which leads to fine-grained loss between adjacent frames as the depth layers increase. Most recently, follow-up works tend to improve this issue in two ways: establish a pure Transformer model [44] or refine temporal convolutions [45]. Although these efforts seem favorable, the robustness remains to be verified due to the lack of open sources. Therefore, our approach preserves the backbone and excavates promising performance with extensive optimization. Particularly, considering the resemblance among CPR actions, we enjoy a specific action segmentation refinement framework [46], laying the foundation for our endeavors to alleviate over-segmentation errors highlighted in the TAS task.

2.3. Prompt-based learning

Served as an evolutionary group of Machine Learning (ML) model training approaches, prompt-based learning preliminarily allows people linguistically specify a certain task for the pre-trained Large Language Model (LLM) to compile and complete [33]. To present a more intuitive perception of prompt-based learning, this section primarily introduces the identification of prompt and prompt engineering. Essentially, a prompt is an instruction depicted in natural language for the model to execute. The procedure of building the ideal prompt for a specific task is called prompt engineering. Subsequently, prompt-based learning involves training a language model on the converted prompt-based dataset. The essence of prompt-based learning is to modify the input into prompts and embed the anticipated output as unfilled blanks within the prompt.

To investigate the effects of prompt-based design in visual applications, models like CLIP [34] and ALIGN [47] have achieved remarkable performance on image recognition tasks. They formulate the objectives as descriptive texts and transit the classification problem into video-text matching. Besides, ActionCLIP [48] proposes a prompt-based paradigm specific to action recognition tasks, aiming to recognize single actions in short video clips. These methods demonstrate the potential of prompt-based learning to motivate the development of visual tasks with multi-modal feature representation.

To this end, we attempt to rethink the effectiveness of frame-wise feature extraction in the TAS task, for which previous action segmentation methods uniformly utilize the pre-trained I3D [30] model. Although the I3D-based features maintain advantageous capacity due to integrating information from the RGB stream and optical-flow stream, they might not be sufficient to construct representative embeddings experiencing complex scenarios, particularly for CPR actions. One closely related work [49] that arose recently, called Br-Prompt, is of significant importance for instructing the feature extraction model equipped with a prompt-based paradigm. On top of that, our approach extends more details elaborated for effective feature extraction of CPR actions in favor of downstream tasks.

3. Methods

In this section, we introduce the proposed Prompt-enhanced hierarchical Transformer (PhiTrans) with clear motivation and objective. Then we distill its integral modules.

3.1. Overall pipeline

As declared in Section 1, our motivation is to investigate *how better could we elevate CPR instruction with action segmentation*, which nails down our objective of devising an exact architecture for CPR action segmentation to excavate its positive effects on CPR instruction.

The overall pipeline of our approach is illustrated in Fig. 1. We first apply Video Features Extractor (VFE) to extract frame-wise features representing potentially involved CPR actions and transition semantics. After that, Action Segmentation Executor (ASE) generates an initial prediction for the input video, explicitly identifying ordinal CPR actions. Finally, to effectively alleviate over-segmentation issues that might give rise to the misclassification of CPR actions, we adopt Prediction Refinement Calibrator (PRC) to refine the performance, thereby producing a final assessment enclosing less boundary ambiguity.

We argue that the aforementioned modules are indispensable, separately playing distinct roles and ultimately bringing down-to-earth performance to the intact model. To support our claims, more module details and corresponding impacts are described below.

3.2. Video Features Extractor

The Video Features Extractor (VFE) module is responsible for extracting discriminative frame-wise features of CPR videos, which subsequently become the input of the action segmentation module. As demonstrated at the bottom of Fig. 1, raw video frames are primarily sampled into neat clips with fixed lengths for local recognition and training efficiency. On top of retaining pre-established semantics by prompt engineering, the VFE module implements visual-linguistic contrastive learning to fuse multi-modal knowledge. Ternary Transformer encoders unify the framework and the synchronous training finally empowers the vision Transformer encoder to produce representative frame-wise features for input videos. The module is inspired by previous methods [48,49], and is adapted to facilitate the feature extraction of CPR actions with more experimental analysis.

3.2.1. Sampling strategy

To unify the inputs and mitigate the footprint burden, we implement a sampling strategy for the raw CPR videos to generate a series of video clips. Specifically, a frame-wise sliding window approach is carried out with the downsampling rate (ds) of frames in each window and the overlapping rate (ol) between windows. The configuration of ds and ol will be illuminated in Section 5.4. The sampling strategy allows the VFE module to concentrate on local dependencies with a fixed video length. Each generated video clip involves one or multiple CPR actions, leading to diverse receptive fields to feature extraction with training efficiency. Moreover, this strategy is vital for augmenting data and empowers the robustness of the VFE module.

3.2.2. Prompt engineering

In order to acquire informative semantic merits for a CPR video clip, prompt engineering resorts to the idea of embedding the expected output string into the input template in a cloze test-like form. We simulate the implementation inspired by Br-Prompt that is non-discriminatorily applied to TAS tasks. In particular, we establish 15 advantageous semantics for the CPR actions appeared in the custom dataset. Therefore, four types of prompts are available to record properties of CPR behaviors such as location, quantity, semantics, and integrality. More specifically, The ordinal prompt z_{ord}^i adopts the format as "*this is the {ith} action in the video*" to simply captures the position of each state. The statistical prompt z_{stat} counts the number of CPR actions of a sequence with a neat format as "*this video clip contains {number of CPR behaviors} actions in total*". To investigate pronounced semantic analysis, we utilize the format "*{ith}, the person is performing the action step of {a certain CPR behavior}*" as the semantic prompt z_{sem}^i focusing
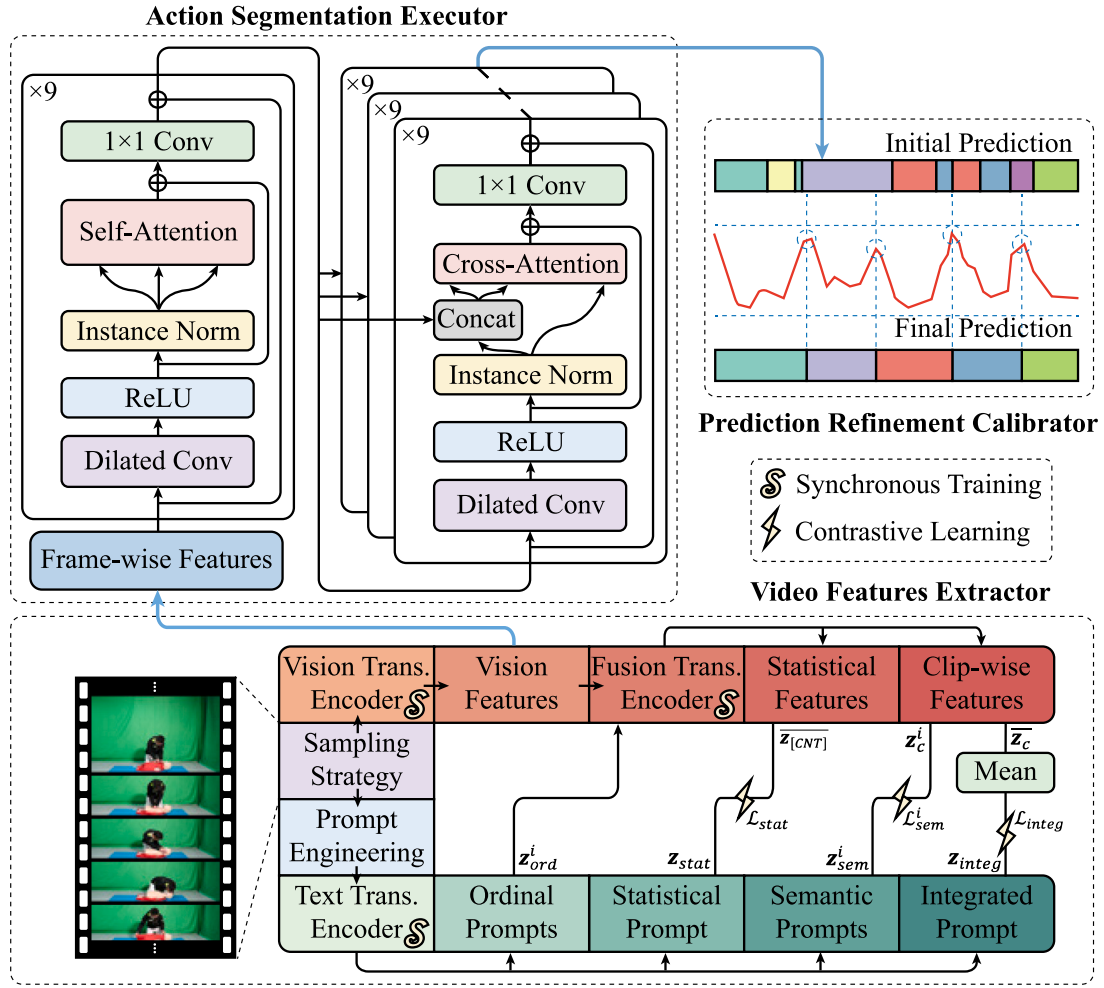
Fig. 1. Overview of PhiTrans pipeline for cardiopulmonary resuscitation action segmentation. The VFE module at the bottom plays a role in generating frame-wise features, which will serve as the input to the top left ASE module to achieve the initial predictions. The PRC module in the top right further implements calibration to yield the eventual predictions. $\mathcal{L}_{stat}$, $\mathcal{L}_{sem}^i$, and $\mathcal{L}_{integ}$ in the VFE module are three elaborate contrastive learning losses. Vision Trans. Encoder, Text Trans. Encoder, and Fusion Trans. Module are synchronously trained. Diverse colors in the PRC module represent action categories, and their length indicates the duration of the corresponding action.

on adjacence and interspersion of actions. Eventually, we regard aggregated semantic prompts in a CPR video clip as the integrated prompt z_{integ}, representing global semantic information. The motivation of prompt engineering aims to learn inter- and intra- affinities of behaviors within CPR videos.

3.2.3. Visual-linguistic contrastive learning

Given massive prompts by the prompt engineering, the VFE module ponders fabulous representation of the CPR video clips through visual-linguistic contrastive learning. To be legible, video clip c and its text description t are introduced to a visual encoder and a text encoder to obtain the corresponding representation z_c and z_t, respectively, and the cosine similarity between the two is expressed as:

$$s(z_c, z_t) = \frac{z_c \cdot z_t}{|z_c||z_t|} \tag{1}$$

The batch similarity matrix S for the video clip feature $\mathcal{Z}_c$ and the text feature $\mathcal{Z}_t$ with batch size B is:

$$S(\mathcal{Z}_c, \mathcal{Z}_t) = \begin{bmatrix} s(z_{c_1}, z_{t_1}) & \cdots & s(z_{c_1}, z_{t_B}) \\ \vdots & \ddots & \vdots \\ s(z_{c_B}, z_{t_1}) & \cdots & s(z_{c_B}, z_{t_B}) \end{bmatrix} \tag{2}$$

We define the ground-truth batch similarity matrix GT, where the similarity of the correct pair is set to 1, and contrastively the error pair is set to 0. The objective of learning is to maximize the similarity between S and GT.

Here we adopt the KL divergence (Kullback–Leibler divergence) as the contrastive loss. For instance, $N \times N$ matrices P and Q, its brief form is defined:

$$D_{KL}(P \parallel Q) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{3}$$

In this way, given dual modal similarity matrix S_C and S_T, we leverage the visual-linguistic contrastive loss as:

$$\mathcal{L} = \frac{1}{2}[D_{KL}(S_C \parallel GT) + D_{KL}(S_T \parallel GT)] \tag{4}$$

To explain comprehensive details while avoiding ambiguous perplexity, we first employ a semantic loss $\mathcal{L}_{sem}^i$, in which visual features z_c^i involving ordinal contents of the video clip contend against commensurate textual semantic representation z_{sem}^i, thereby maintaining substantial semantic concernments. Meanwhile, the average pooling video clip features $\overline{z_c}$ containing ordinal collections compete with integrated prompts representation z_{integ} via the integrated loss $\mathcal{L}_{integ}$ to allow holistic demonstration. Furthermore, to transparently exert a profound influence on the quantitative information, a statistical loss $\mathcal{L}_{stat}$ is applied by way of contrasting average pooling enumeration features $\overline{z_{[CNT]}}$ with statistical prompts representation z_{stat}. We ablate these loss components in Section 5.6.2.

Finally, the united loss of the model is manifested as:

$$\mathcal{L} = \sum_{i=1}^{K} \mathcal{L}_{sem}^i + \mathcal{L}_{integ} + \mathcal{L}_{stat} \tag{5}$$

3.3. Action Segmentation Executor

The Action Segmentation Executor (ASE) module takes frame-wise features derived from the VFE module as input and produces moderate CPR action segmentation predictions. The structure of the ASE module is illustrated in the upper left of Fig. 1. Generally, the ensemble is a hierarchical encoder–decoder architecture, wherein the output of the encoder as initial predictions will flow in ternary decoders for distillation. At the same time, the ASE module attempts to acquire hierarchical perception with window-perceptive self-attention and dilated temporal convolution throughout the framework.

3.3.1. Hierarchical encoder–decoder

ASE module is composed of one encoder and three identical decoders. Each of them contains nine blocks to provide hierarchical representation. After generating the initial prediction by the encoder, three decoders perform a refinement process to boost the performance. The input to the encoder is a sequence of pre-extracted frame-wise features in sizes $D \times S$, where D refers to the feature dimension of each frame and S represents the total number of frames of the input video. The first layer of the encoder uses a fully connected layer to adjust the input feature dimension. Subsequently, each encoder block utilizes a dilated temporal convolution as a feed-forward layer, which is followed by a ReLU activation function and instance normalization, connecting the single-head self-attention layer. A residual concatenation is taken between the two layers, and then the output dimension is reshaped by 1×1 convolution to join the next encoder block. The final encoder block outputs the initial prediction $p_e \in R^{S \times C}$ by passing a fully connected layer, where C represents the number of action categories. Afterward, the decoder takes the initial prediction as input and is arranged in accordance with the encoder as a whole, except for the cross-attention layer. The cross-attention layer combines the results from the encoder with the output from the previous layer, treating the aggregation as the query Q and the key K, and the output from the previous layer as the value V. The advantage of this manner is that the frame-wise confidence scores from the encoder can be involved in the refinement stage by generating attention weights, and these attention weights are utilized for linear concatenation, without affecting the feature space V itself. In the end, the model makes the best use of three identical decoders to hierarchically produce the final prediction.

3.3.2. Hierarchical perception

Two hierarchical strategies are simultaneously adopted to scratch multi-scale receptive fields by enlarging the window size of the self-attention layer and dilated temporal convolution throughout the ASE module. Considering videos in the CPR dataset tend to cover thousands of frames, it is fairly demanding to seize significant vision plots for the self-attention layer within each block of the encoder or the decoders. We follow the spirit of hierarchical representation patterns proposed in [43] to mitigate this issue. Concretely, such a strategy first concentrates on the local semantics and then gradually enlarges the receptive field to acquire the global concepts, allowing the model to learn extrinsic-to-intrinsic knowledge of CPR actions displayed in the videos. In addition, both apparent differences and subtle discrepancies are taken into account, yielding a more specific fashion to the traits of CPR actions. Practically, the window-perceptive self-attention layer calculates attention weights with each particular frame within its local window at a w scale, which is doubled as the blocks stack deeper (i.e., $w = 2^i$, $i = 1, 2, \ldots$). Similarly, to introduce constructive local inductive bias, we follow the dilated temporal convolution as utilized in the TAS task by expanding the kernel size consistent with the self-attention layer.

3.4. Prediction Refinement Calibrator

The Prediction Refinement Calibrator (PRC) module intends to remarkably alleviate over-segmentation errors of the predictions from the ASE module. The PRC module allies with and refines the ASE module by duplicating the segmentation pipeline but replacing the objective from action predictions with boundary probabilities. In particular, a comprehensive loss function serves as the motivation for jointly training the ASE and PRC modules. The overall loss function is a combination of the action segmentation loss and the boundary regression loss, with correlative ablation conducted in Section 5.6.2. As shown in the upper right of Fig. 1, the generated boundary probability curves further calibrate the action predictions from the ASE module with less over-segmentation issues and present the final CPR action segmentation predictions.

3.4.1. Boundary probability calibration

To clarify the assets of the PRC module, we depict its intrinsic concept for boundary probability calibration. During the inference period, the PRC module regresses frame-wise action boundary possibilities $P \in [0, 1]^S$. Subsequently, the action boundaries $B \in \{0, 1\}^S$ are determined by electing multiple P_s from P as the local maximum and simultaneously fulfilling the exceeding condition of the threshold $p = 0.5$.

In practice, the ASE module primarily generates initial predictions by assigning the action categories to the action segments with potential over-segmentation issues. Then the calculated action boundaries B compartmentalize the video clip into refined action segments, each of which contains only one action both theoretically and practically. The retained action categories by majority voting for each action segment reach the final prediction of the entire model.

3.4.2. Action segmentation loss

The overall action segmentation loss is defined as:

$$\mathcal{L}_{as} = \frac{1}{G} \sum_g \left(\mathcal{L}_{cls}^* + \mathcal{L}_{smo} \right) \tag{6}$$

More specifically, the classification loss $\mathcal{L}_{cls}^*$ utilizes median frequency balancing, where the action weights of each action category in the temporal action segmentation task are calculated by dividing the mean frequency of each action category by the frequency of each action class. Concerning the smoothing loss $\mathcal{L}_{smo}$, we implement Gaussian Similarity-weighted Truncated Mean Squared Error (GS-TMSE), which penalizes all frames in a video clip for the purpose of smoothing the transition of action probabilities between frames while preventing them from interfering with the frames where actions de facto transition. Its concrete form is as follows:

$$\mathcal{L}_{GS-TMSE} = \frac{1}{SC} \sum_{s,c} \exp\left(-\frac{\|\boldsymbol{x}_s - \boldsymbol{x}_{s-1}\|^2}{2\sigma^2} \right) \delta_{s,c}^2 \tag{7}$$

$$\delta_{s,c} = \min\{ |\log p_{s,c} - \log p_{s-1,c}|, \ \tau \} \tag{8}$$

where $\boldsymbol{x}_s$ is the similarity index of the frame s, σ denotes the variance and is simply set to 1, and the threshold τ is set to 4. The advantageous implementation benefits from the property of GS-TMSE. For brevity, the Gaussian kernel based on the similarity of frames (frame-wise features in our experiments), the function punishes contiguous but discriminative frames with merely a small weight. We average the losses of G stages (in our framework $G = 4$) as the overall action segmentation loss.
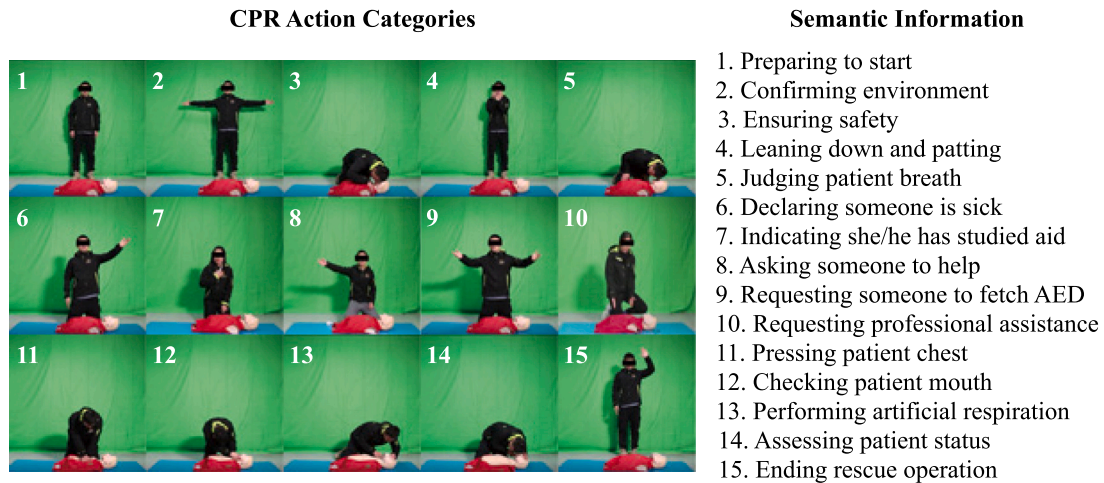
CPR Action Categories

Semantic Information

1. Preparing to start
2. Confirming environment
3. Ensuring safety
4. Leaning down and patting
5. Judging patient breath
6. Declaring someone is sick
7. Indicating she/he has studied aid
8. Asking someone to help
9. Requesting someone to fetch AED
10. Requesting professional assistance
11. Pressing patient chest
12. Checking patient mouth
13. Performing artificial respiration
14. Assessing patient status
15. Ending rescue operation

Fig. 2. Visualization of 15 action categories recorded in the CPR dataset and their corresponding semantic information.

Table 1
Summary of datasets for model development and application.

Dataset	Videos	Classes	Duration	fps	Instances	View
50Salads	50	17	6.4	30	20	Top
Breakfast	1712	48	≈1	15	6	Third person
GTEA	28	11	≈1	15	20	Egocentric
CPR	**99**	**15**	**2**	**15**	**17**	**Third person**

3.4.3. Boundary regression loss

We employ a weighted logistic regression loss function to endow our model with boundary-aware capacity, which is defined as:

$$\mathcal{L}_{br} = \frac{1}{SG} \sum_{g} \sum_{s=1}^{S} \left(w_p y_s \cdot \log p_s + (1 - y_s) \cdot \log(1 - p_s) \right) \tag{9}$$

where y_s and p_s represent the ground-truth action boundary labels and predicted action boundary possibilities for the frame s, respectively. Since the number of action boundary frames is much smaller than that of other frames, the factor w_p is kindly devised to samples with weight positive. To further clarify w_p, it comes from the reciprocal ratio of positive data points over the entire training data.

4. Dataset

We primarily compare the performance of various models on three challenging public datasets that are widely recognized in the TAS task concentrating on catering behaviors, whereby the optimal architecture will be established especially for CPR action segmentation. We argue for the reasonability of selecting these three datasets on account of analogous representation patterns compared to the CPR actions, wherein the behaviors are globally invariant and locally distinct. More importantly, we introduce the preprocesses of the custom CPR dataset and explain its implication. The essential takeaway information for these four datasets is summarized in Table 1.

4.1. Preliminary public datasets

50Salads dataset [50] is beneficial for research in action recognition, activity detection, process tracking, etc., which collects 50 videos from 25 participants preparing two different kinds of mixed salads, and contains more than 4 h of video data. Each video lasts 6.4 min long on average and consists of about 20 action instances. All videos are recorded from a top-down view, including 17 action category annotations.

Breakfast dataset [51] is related to the preparation of daily breakfast. It comprises 18 different kitchen scenes with diverse backgrounds conducive to monitoring and analyzing daily human activities. The dataset contains 1712 videos with 48 action annotations recorded from a third-person view of 52 participants, wherein each video represents an average of 6 action instances.

GTEA dataset [52] contains 28 egocentric videos with 11 action categories involving daily kitchen activities performed by 4 participants. On average, each video has 20 action instances and is about a half-minute long.

4.2. Custom CPR dataset

We collect the CPR dataset with ethical approval and notification of subjects with the assistance of MIRAGESTARS Inc., which involves 99 videos of participants performing the whole process of cardiopulmonary resuscitation in a standard green screen laboratory environment. This work is conducted in adherence to the tenets of the Declaration of Helsinki, and ethics approval is obtained by the Ethics Committee of the Shenzhen International Graduate School of Tsinghua University Submission F111/2022.

Each video has about two minutes duration on average, containingup to 15 action categories shown in Fig. 2 with their representative frames and the corresponding semantic information. The raw videos shot are identically transformed utilizing the stream processing tool called FFmpeg. More concretely, the frame rate and resolution of videos are decreased from 25 fps, 2k to 15 fps, 720 p respectively, on account of alleviating the calculating burden of proposed models, while ensuring the accessibility of the pipeline and the efficiency of the architecture to the maximum extent. It is worth noting that the simplifications above significantly reduce the video size from 320 MB–1.1 GB to 1.9 MB–5.3 MB, followed by wiping out the audio channel. These transformations sufficiently consider that CPR action segmentation generally depends on integral gestures rather than pixel-wise identification, whereby hardly any practical loss will be generated. After that, these videos are labeled at a frame level by experienced trainers. The crafted CPR dataset possesses diverse challenges, including but not limited to the transience of partial actions, resemblance among actions, and out-of-place actions.

The motivation of this dataset is prone to assist CPR instruction by automatically identifying potential omission, repetition, or out-of-order situations at a second-wise level, even if taking frame-wise misregistration into consideration. Furthermore, we adopt to implement four-fold nested cross validation to minimize latent optimistically biased evaluation, thus fairly revealing the performance of the presented model.

Table 2

Action segmentation performance of various state-of-the-art (SOTA) models on GTEA, 50Salads datasets and Breakfast.

Dataset	GTEA					50Salads					Breakfast				
Model	F1@{10,25,50}			Edit	Acc.	F1@{10,25,50}			Edit	Acc.	F1@{10,25,50}			Edit	Acc.
MS-TCN [24]	85.8	83.4	69.8	79.0	76.3	76.3	74.0	64.5	67.9	80.7	52.6	48.1	37.9	61.7	66.3
DTGRM [28]	87.8	86.6	72.9	83.0	77.6	79.1	75.9	66.1	72.0	80.0	68.7	61.9	46.6	68.9	68.3
BCN [23]	88.5	87.1	77.3	84.4	_79.8_	82.3	81.3	74.0	74.3	84.4	68.7	65.5	55.0	66.2	_70.4_
MS-TCN++ [42]	88.8	85.7	76.0	83.5	80.1	80.7	78.5	70.1	74.3	83.7	64.1	58.6	45.9	65.6	67.6
ASRF [46]	89.4	87.8	_79.8_	83.7	77.3	84.9	_83.5_	_77.3_	79.3	84.5	74.3	68.9	56.1	72.4	67.6
SSTDA [39]	90.0	_89.1_	78.0	86.2	_79.8_	83.0	81.5	73.8	75.8	83.2	75.0	69.1	55.2	73.7	70.2
UVAST+alignment decoder [40]	77.1	69.7	54.2	_90.5_	62.2	_86.2_	81.2	70.4	**83.9**	79.5	**76.7**	_70.0_	_56.6_	**77.2**	68.2
ASFormer [43]	_90.1_	88.8	79.2	84.6	79.7	85.1	83.4	76.0	79.6	_85.6_	76.0	**70.6**	**57.4**	_75.0_	**73.5**
Br-Prompt [49]+ASFormer	**94.1**	**92.0**	**83.0**	**91.6**	**81.2**	**89.2**	**87.8**	**81.3**	_83.8_	**88.1**	N/A	N/A	N/A	N/A	N/A

5. Experiments and results

5.1. Evaluation metrics

5.1.1. Frame-wise accuracy (Acc.)

Frame-wise accuracy is commonly used as an evaluation metric for action segmentation [50,51,53], whereas it is easily affected by long-duration actions and not sensitive to over-segmentation issues.

5.1.2. Segmental edit score (Edit)

Segmental edit score [54] is used to assess the model performance in predicting the ordering of action segmentation without being affected by minor temporal shifts. Once proposed, the segmental edit score has been widely used in many temporal action segmentation tasks [23,28,29,39,53] since it combines the assessment of accuracy and efficiency into a single metric. There is considerable uncertainty about when one action will cease and another will begin. Typically, in practical applications such as surgical workflow assessment, the accurate temporal continuity of surgical operations tends to be more crucial than precise temporal segmentation, as the same goes for CPR instruction.

5.1.3. Segmental F1 score with overlapping threshold k (F1@k)

Segmental overlap $F1$ score [38] has three distinctive characteristics: (1) penalizes over-segmentation errors; (2) ignores minor temporal shifts between the predictions and ground-truth; (3) is determined by the total number of actions but does not depend on the duration of each action instance. By comparing the Intersection over Union (IoU) score of the predictions and ground-truth, if the threshold $\tau = \frac{k}{100}$ is exceeded, it is determined as true positive, otherwise as true negative, where $k = 10, 25, 50$ are adopted in temporal action segmentation tasks. The $precision$ and $recall$ are defined as $precision = \frac{true\ positives}{true\ positives + false\ positives}$, $recall = \frac{true\ positives}{true\ positives + false\ negatives}$, respectively. Then $F1@k$ value can be computed from $F1 = 2 \times \frac{precision*recall}{precision+recall}$.

5.2. Implementation details

We train the VFE module for 50 epochs using AdamW [55] optimizer with a learning rate of 5e-6 and weight decay of 0.2. There is a warm-up in the first 5 epochs. The batch size is 12. The downsampling rate and overlapping rate of the VFE module are experimentally determined in Section 5.4, which are 4, 8, 12 and 2, 1, 1, respectively. We adopt a spatial resolution of 224 × 224 as input and the feature dimension of output for each frame is 768. The vision and text Transformer encoders are synchronously pre-trained on Kinetics-400 [30] derived from [48]. Following the standard setups from previous methods [24, 42], we train the ASE and PRC module for 50 epochs using Adam optimizer with a learning rate of 5e-4 and batch size of 1. We follow the same hyperparameter settings as per [43,49]. For evaluation, we perform 4-fold cross validation and report the average results. The VFE module is trained in identical split configurations to avoid information leakage. All experiments are implemented with PyTorch 1.12 on 4 NVIDIA GeForce RTX 3090 GPUs.

Table 3

Action segmentation performance with features following various window configurations on the CPR dataset (#split 1).

Window configuration	F1@{10,25,50}			Edit	Acc.
Baseline	95.8	93.6	89.8	94.2	89.5
$ds = [2, 4]$ & $ol = [2, 2]$	95.3	94.8	91.5	93.3	**92.0**
$ds = [2, 4, 8]$ & $ol = [4, 2, 1]$	**96.5**	95.5	92.0	94.8	91.0
$ds = [4, 8, 12]$ & $ol = [2, 1, 1]$	96.0	**95.7**	**92.7**	**95.0**	91.8

5.3. Module selection for action segmentation

It is non-trivial to select an appropriate and distinctive backbone to approach our goal of CPR action segmentation. Table 2 reveals the best temporal action segmentation models on GTEA, 50Salads, and Breakfast datasets. Comparing the performance on public datasets is intuitively convincing, for which we gaze at the optimal one, AS-Former [43]. Particularly, we overlook Br-Prompt [49] in this subsection, which plays a role in feature representation, which enhances the action segmentation models rather than serves as one of them.

Considering the property of the collected CPR dataset, we lay more emphasis on the analysis capacity to confront more enormous and complex data. Concretely, on the GTEA and 50Salads datasets with slightly smaller video volumes, we find ASFormer exhibits inference accuracies comparable to ASRF [46] and SSTDA [39]. While carried out on the Breakfast dataset with a multitude of videos, the model explicitly yields state-of-the-art performance. Through observing the performance on three public datasets, ASFormer illustrates the application of Transformer in the TAS task, affirming the non-negligible capacity of temporal representation and persistent sensitivity of long-term relationships. Due to the convenient scalability and robustness of ASFormer, we adopt this model as a vanilla backbone for CPR action segmentation.

5.4. Sampling strategy selection for feature extraction

In the process of extracting video frame-wise features, we adopt a specific sampling strategy dependent on pertinent experiments to generate video clips of a fixed length, which serve as the input of the video features extractor. In general, each video clip adopts a 16-frame window. We are mainly concerned with the downsampling rate (ds) of frames in each window, and the overlapping rate (ol) between two windows. It should be emphasized that both the downsampling rate and the overlapping rate are heuristically selected according to the characteristics of the dataset. Concretely, longer windows lead to sparse information about each action, while shorter ones are even unable to contain more than two actions. We empirically take both long and short windows into consideration, which produce excellent feature extraction performance by involving multi-scale information.

Diverse window configurations are displayed in Table 3, which indicate the non-trivial importance of the downsampling rate and the overlapping rate. The collective results are retained on the custom CPR dataset (#split 1) under the support of the ASE module as the action
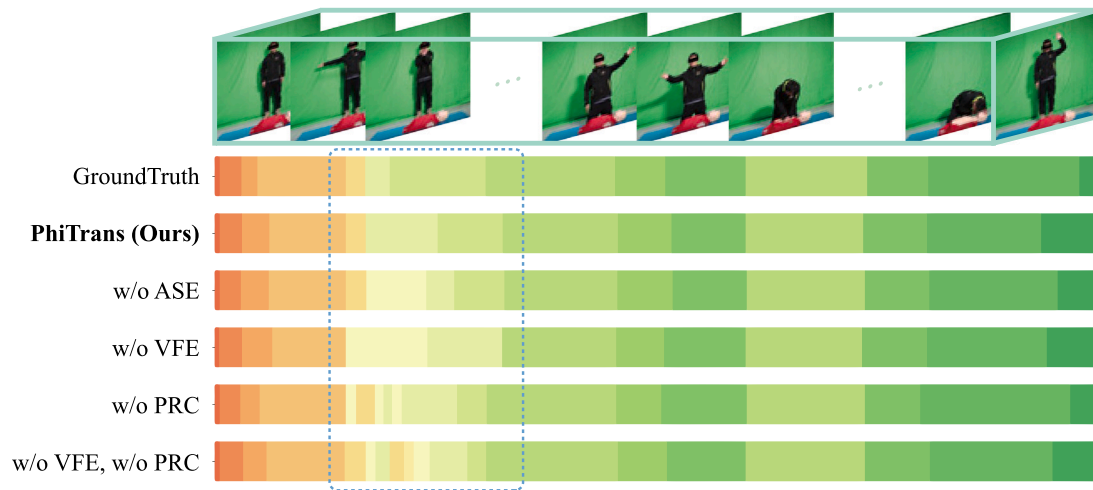
Fig. 3. Visualization of PhiTrans and related variants for temporal action segmentation on the CPR dataset. Diverse colors represent action categories, and their length indicates the duration of the corresponding action.

Table 4
Model performance compared with baselines and competitive models on the CPR dataset.

Model	#params.	FLOPs	F1@{10,25,50}			Edit	Acc.
MS-TCN [24]	0.8M	1.6G	93.3	91.6	85.5	89.9	88.7
MS-TCN[a]	0.7M	1.4G	94.0	92.4	88.7	91.8	90.0
MS-TCN++ [42]	1.0M	2.0G	93.6	92.1	87.1	90.5	88.9
MS-TCN++[a]	0.9M	1.8G	93.8	92.7	88.6	91.2	90.3
UVAST+alignment decoder [40]	2.2M	3.6G	95.3	93.7	89.0	93.6	89.3
UVAST+alignment decoder[a]	1.7M	2.3G	**96.0**	94.4	90.6	93.9	90.2
PhiTrans (Ours)	1.5M	3.1G	95.8	**95.3**	**91.7**	**94.6**	**91.1**

[a] Indicates that features used are extracted from the VFE module of PhiTrans.

segmentation backbone. Multiple downsampling rates and overlapping rates are adopted simultaneously as mentioned above. In addition, training efficiency is an essential factor to be carefully considered. Complicated windows conduce to substantial training time, while simple ones may fail to grab discriminative information. In the light of this, we conduct three groups of window arrangements to seek out the optimal one. Here we adopt the preceding Transformer-based method [43] as the vanilla baseline.

When the downsampling rates are 2, 4, and correspondingly the overlapping rates are 2, 2, redundant local features are captured. Each window contains only few actions, where a good deal of frames belong to the same action, resulting in poor effects. Similarly, employing 2, 4, and 8 downsampling rates and 4, 2, and 1 overlapping rates though reaches a high *F1@50* as 96.5, its overlapping rate of 4 misleads the module extensively reusing the same information, decreasing the abundance of video features. We finally adopt downsampling rates as 4, 8, and 12 corresponding to the overlapping rates of 2, 1, and 1, which balance the expressions derived from various receptive fields with rational training efficiency.

5.5. Model performance of PhiTrans

5.5.1. Quantitative results

Our model accomplishes the objective to serve as a productive toolbox assisting CPR instruction with action segmentation, with the intact model performance manifested in Table 4. On the whole, our model is adequate to approach the challenge of assisting CPR instruction, performing well on all metrics used surpassing 91.0%. The comprehensive performance reveals the effectiveness and robustness of PhiTrans. More importantly, PhiTrans reaches 94.6% on *Edit*, a metric which reflects the model ordinal predicting performance wiping out fine-drawn temporal shifts. That means our model is prone to understand the chronological relationship of CPR actions that practically

necessitates. In addition, 95.8%, 95.3%, and 91.7% on *F1@{10,25,50}* indicate that PhiTrans successfully penalizes over-segmentation errors highlighted in the TAS task. Our model significantly outperforms two typical baselines [24,42] and also defeats the competitive model [40] on the CPR dataset. Furthermore, we equip these models with the features extracted from our VFE module to demonstrate the salience of both the feature extraction and the segmentation of our model. Specifically, though all three models benefit from the features derived from our VFE module, i.e., gain of all metrics, release of parameters and computational complexity, they fall behind PhiTrans to varying degrees. Though not designed to pursue a state-of-the-art performance on the public dataset, PhiTrans outperforms cutting-edge models without any fine-tuning, revealed in Table 6. It gains +0.8% improvement on *F1@50*, with other metrics comparable to previous state-of-the-art, which further manifests the effectiveness of PhiTrans on the general TAS task.

5.5.2. Qualitative results

To improve the model interpretability for empirical observation and analysis, we visually present the favorable performance of PhiTrans and its modules. Primarily, Fig. 6 visualizes the confusion matrix of PhiTrans on the CPR dataset, which witnesses dominant segmentation performance on the majority of CPR actions by PhiTrans. Although our model exhibits hesitation in distinguishing actions ranging from 6 (*declaring someone is sick*) to 10 (*requesting professional assistance*), it is rational since these actions primarily differ in terms of vocal expression by the subjects, which is not incorporated into our modality for accommodating complex real-world scenarios. Besides, we show the distributions of features generated by the VFE module in Fig. 4. Distinct clusters imply good feature representation, and the potential overlap denotes that some CPR actions are similar and interweaved. To indicate that the ASE and PRC modules play essential roles, Fig. 5
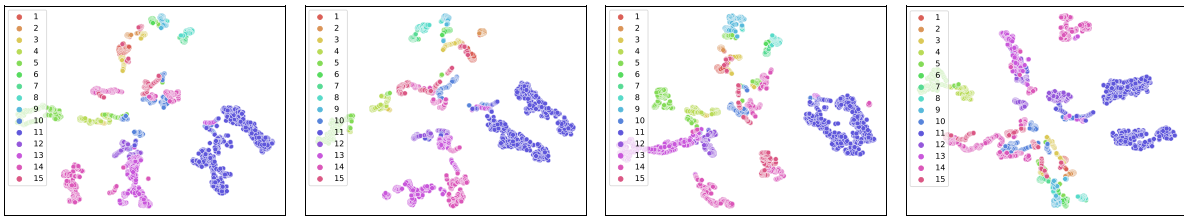
Fig. 4. Distributions of features generated by the VFE module visualized by t-SNE [56]. The video samples are randomly selected from each of the four-fold cross-validation splits of the CPR dataset for feature presentation. Points denote the frame features in the video, and diverse colors represent action categories.
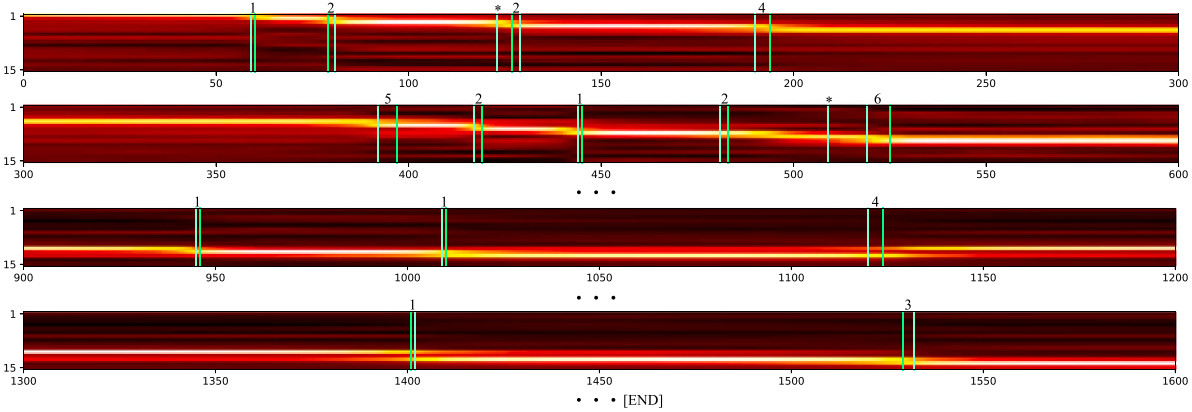

Fig. 5. Cross-attention map visualization of PhiTrans on the CPR dataset. The horizontal and vertical axes indicate the index of the frames of the video and action categories, respectively. The brighter color implies a higher attention score. The green and blue lines indicate positions of ground-truth and refined prediction boundaries, respectively. The numbers above two lines measure their distances in frames. ∗ presents the occurrence of over-segmentation.
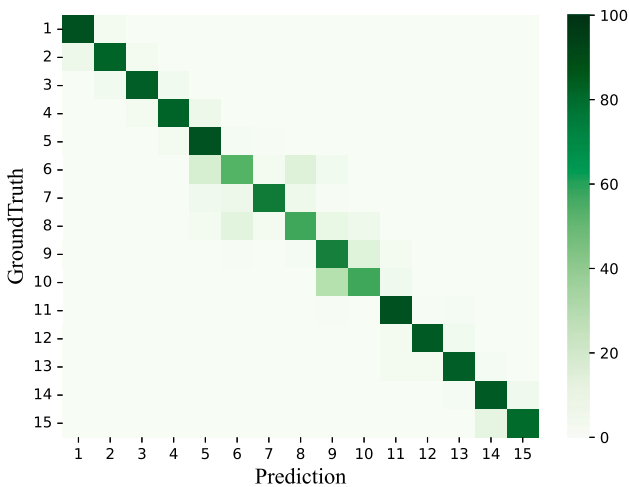

Fig. 6. Confusion matrix for the custom CPR dataset. The numbers on the horizontal and vertical axes represent the ordinal CPR actions. The numerical unit in the confusion matrix is expressed in percentages.

Table 5
Module ablation of PhiTrans on the CPR dataset.

Model with variants	Components			F1@{10,25,50}			Edit	Acc.
	ASE	VFE	PRC					
All	✓	✓	✓	**95.8**	**95.3**	**91.7**	**94.6**	**91.1**
w/o ASE	×	✓	✓	94.3	93.0	88.5	92.9	88.6
w/o VFE	✓	×	✓	94.1	91.7	87.1	91.8	87.6
w/o PRC	✓	✓	×	95.3	94.5	91.5	93.6	91.1
w/o VFE, w/o PRC	✓	×	×	95.3	94.2	89.5	93.4	89.7

5.6. Ablation study of PhiTrans

5.6.1. Module ablation

Table 5 quantitatively reveals the effect of each module on holistic model performance. We detach each part independently to observe any drop. We experiment our model without the ASE module, w/o ASE, by replacing the ASE module with the approved backbone [24], w/o VFE by adopting ResNet152 backbone-based frame-wise features as input, and w/o PRC by merely applying loss functions proposed in previous methods [43]. The results show that lacking any module will lead to an apparent performance drop on all metrics, manifesting the essentiality of each module. Worst of all, the loss of the VFE module (w/o VFE) causes the model performance to reach a trough, leading to a disastrous result of the total −9.9% on *F1* and −2.8%, −3.5% on *Edit* and *Acc.*, respectively. Interestingly, the further removal of the PRC module on this basis (namely, w/o VFE, w/o PRC) instead presents a relative increase on all metrics. This situation can be explained by the fact that the PRC module tends to suppress over-segmentation issues excessively when the feature expression is insufficient. These results prove the indispensability of the VFE module for its splendid feature representation. Similarly, whichever module misses accounts for the loss of corresponding capacity.

Besides, to intuitively present effects of these modules, Fig. 3 visualizes the qualitative representation on the custom CPR dataset. Under

visualizes the cross-attention map of PhiTrans on the CPR dataset. We observe that the ASE module can handle both the inter- and intra-conditions with consistently high scores for single action and distinct transitions between different action categories. The PRC module makes the predictions closer to the ground-truth by calibrating the boundaries. Although a few over-segmentation issues still exist as lines with the signal ∗ in the figure, most predicted boundaries are only a few frames away from the ground-truth. These results qualitatively illustrate the superiority of our model.

Table 6
Model performance of PhiTrans on the 50Salads dataset.

50Salads	F1@{10,25,50}			Edit	Acc.
Cutting-edge [49]	89.2	87.8	81.3	**83.8**	88.1
PhiTrans (ours)	**89.3**	**87.8**	**82.1**	83.4	**88.1**

Table 7
Loss ablation of PhiTrans on the CPR dataset (#split 1).

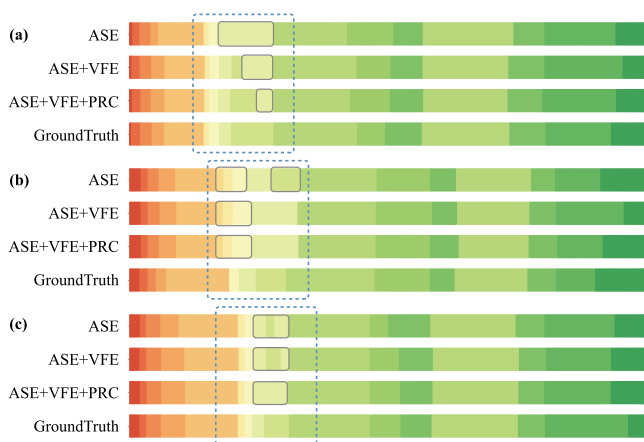| VFE loss components	F1@{10,25,50}			Edit	Acc.
$\mathcal{L}_{sem}$	94.1	91.6	88.1	92.7	88.7
$\mathcal{L}_{sem} + \mathcal{L}_{integ}$	95.1	93.9	90.1	92.8	90.0
$\mathcal{L}_{sem} + \mathcal{L}_{integ} + \mathcal{L}_{stat}$	**96.2**	**95.5**	**91.4**	**95.5**	**90.5**
PRC loss components	F1@{10,25,50}			Edit	Acc.
$\mathcal{L}_{as}$	93.7	92.4	88.3	91.0	89.4
$\mathcal{L}_{as} + \mathcal{L}_{br}$	**96.2**	**95.5**	**91.4**	**95.5**	**90.5**

Fig. 7. Examples of remaining segmentation problems for PhiTrans on the CPR dataset. PhiTrans fails to completely eliminate fragments in the adjacent actions, suffers trivial segments, and loses direction encountering similar actions. Note that diverse colors represent action categories, and their length indicates the duration of the corresponding action.

the reinforcement of each module, PhiTrans could not only perceive the presence of delicate actions but alleviate over-segmentation issues to a great extent, as displayed in the blue dashed line. Specifically, multiple false fragments are significantly denied by employing the PRC module, which can be qualitatively exemplified in the second and fifth rows of Fig. 3. Three hands-on modules proportionally bring promising performance for the challenge of assisting CPR instruction.

5.6.2. Loss ablation

Table 7 presents two types of loss components that arise in the proposed model, one for extracting frame-wise features and the other for refining segmentation. Specifically, all three VFE loss components contribute to capturing representative features. The semantic loss $\mathcal{L}_{sem}$ supervises distinctive features related to various CPR actions, laying the foundation for feature extraction. As the integrated $\mathcal{L}_{integ}$ and statistical loss $\mathcal{L}_{stat}$ are successively added, the model achieves respective improvement, indicating the indispensability of all losses. Additionally, it can be witnessed that both loss components utilized in the PRC module advocate the model performance by action segmentation control and boundary regression refinement. Notably, the boundary regression loss $\mathcal{L}_{br}$ brings + 3.0% on every *F1* and *Edit*, exactly showing its effectiveness to alleviate over-segmentation issues. Overall, these loss ablations certify the necessity and trait of losses implemented in our model.

6. Discussion

Although PhiTrans is capable of efficiently modeling the characteristics of each action when combining the latent information between

actions at various scales, there still inevitably exist segmentation errors from diverse aspects while facing demanding fragments. Specifically, as shown by the dashed line in Fig. 7(a), although the difficult-to-recognize action segments are achieving better accuracy with the consecutive reinforcement of the model, the over-segmentation errors cannot be eliminated entirely, resulting in remaining lingering clips from the adjacent action in the final segmentation result. Furthermore, PhiTrans fails to avoid the redundant segments as shown in Fig. 7(b) due to the short duration of some actions. Finally, considering the introduction of boundary regression loss to alleviate over-segmentation issues, certain contiguous actions have similar loss curve peaks, which may mislead the model into classifying the current action as an adjacent category. Fig. 7(c) demonstrates the over-correction state caused by the above situation. To tackle these remaining problems, an additional modality, voice, can be fused to guide the estimation of visually similar CPR actions mentioned in Section 5.5.2. Besides, we will consider more effective Transformer backbones like Swin Transformers [57,58] and pursue an end-to-end paradigm in the future work.

7. Conclusion

In this study, we originally deliberate a rewarding pipeline that assists the enhancement of CPR instruction via action segmentation through novel deep learning architectures. Specifically, we collect a custom CPR dataset involving videos of the whole process of cardiopulmonary resuscitation assessment along with corresponding frame-level semantic annotation. Accordingly, we devise a Prompt-enhanced hierarchical Transformer, called PhiTrans, especially for CPR action segmentation. PhiTrans consists of three integral modules: Video Features Extractor, Action Segmentation Executor, and Prediction Refinement Calibrator. Such an architecture adequately considers the characteristics of CPR actions, facilitating the capacity to capture transient CPR actions while maintaining impressive performance. Extensive ablation experiments present that PhiTrans inspires improvement compared to half-baked models on the CPR dataset.

In conclusion, it is non-trivial that PhiTrans is committed to CPR action segmentation, which is conducive to freeing experts of detecting and rectifying potential ordinal or oblivious mistakes made by subjects, thereby manifesting a compelling pipeline on elevating CPR instruction with action segmentation. The results of this research may serve as the cornerstone and offer a route towards developing a prospective orientation that would leverage fine-grained criteria like CPR action assessment.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

References

[1] S.C. Brooks, G.R. Clegg, J. Bray, C.D. Deakin, G.D. Perkins, M. Ringh, C.M. Smith, M.S. Link, R.M. Merchant, J. Pezo-Morales, et al., Optimizing outcomes after out-of-hospital cardiac arrest with innovative approaches to public-access defibrillation: A scientific statement from the international liaison committee on resuscitation, Circulation 145 (13) (2022) e776–e801.

[2] C. Sandroni, T. Cronberg, M. Sekhon, Brain injury after cardiac arrest: pathophysiology, treatment, and prognosis, Intensive Care Med. (2021) 1–22.

[3] A.H. Association, et al., Standards for cardiopulmonary resuscitation (cpr) and emergency cardiac care (ecc), JAMA 227 (1974) 837–860.

[4] T. Lancet, Out-of-hospital cardiac arrest: a unique medical emergency, Lancet 391 (2018).

[5] I. Hasselqvist-Ax, G. Riva, J. Herlitz, M. Rosenqvist, J. Hollenberg, P. Nordberg, M. Ringh, M. Jonsson, C. Axelsson, J. Lindqvist, et al., Early cardiopulmonary resuscitation in out-of-hospital cardiac arrest, N. Engl. J. Med. 372 (24) (2015) 2307–2315.

[6] L. Wik, P.A. Steen, N.G. Bircher, Quality of bystander cardiopulmonary resuscitation influences outcome after prehospital cardiac arrest, Resuscitation 28 (3) (1994) 195–203.

[7] E.J. Gallagher, G. Lombardi, P. Gennis, Effectiveness of bystander cardiopulmonary resuscitation and survival following out-of-hospital cardiac arrest, JAMA 274 (24) (1995) 1922–1925.

[8] S. Pivač, P. Gradišek, B. Skela-Savič, The impact of cardiopulmonary resuscitation (cpr) training on schoolchildren and their cpr knowledge, attitudes toward cpr, and willingness to help others and to perform cpr: mixed methods research design, BMC Public Health 20 (2020) 1–11.

[9] M.Y. Khanji, C.A.A. Chahal, F. Ricci, M.W. Akhter, R.S. Patel, Cardiopulmonary resuscitation training to improve out-of-hospital cardiac arrest survival: addressing potential health inequalities, Eur. J. Prev. Cardiol. 29 (17) (2022) 2275–2277.

[10] K. Bielski, B.W. Böttiger, M. Pruc, A. Gasecka, M. Sieminski, M.J. Jaguszewski, J. Smereka, N. Gilis-Malinowska, F.W. Peacock, L. Szarpak, Outcomes of audio-instructed and video-instructed dispatcher-assisted cardiopulmonary resuscitation: a systematic review and meta-analysis, Ann. Med. 54 (1) (2022) 464–471.

[11] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, S. Lacoste-Julien, Unsupervised learning from narrated instruction videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4575–4583.

[12] A. Piergiovanni, A. Angelova, M.S. Ryoo, I. Essa, Unsupervised discovery of actions in instructional videos, 2021, arXiv preprint arXiv:2106.14733.

[13] G. Ding, A. Yao, Temporal action segmentation with high-level complex activity labels, IEEE Trans. Multimed. (2022).

[14] Y. Xu, X. He, G. Xu, G. Qi, K. Yu, L. Yin, P. Yang, Y. Yin, H. Chen, A medical image segmentation method based on multi-dimensional statistical features, Front. Neurosci. 16 (2022) 1009581.

[15] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri, Inf. Fusion 91 (2023) 376–387.

[16] X. He, G. Qi, Z. Zhu, Y. Li, B. Cong, L. Bai, Medical image segmentation method based on multi-feature interaction and fusion over cloud computing, Simul. Model. Pract. Theory 126 (2023) 102769.

[17] Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, Y. Liu, X-net: a dual encoding–decoding method in medical image segmentation, Vis. Comput. (2021) 1–11.

[18] E. Apostolidis, E. Adamantidou, A.I. Metsai, V. Mezaris, I. Patras, Video summarization using deep neural networks: A survey, Proc. IEEE 109 (11) (2021) 1838–1863.

[19] S. Vishwakarma, A. Agrawal, A survey on activity recognition and behavior understanding in video surveillance, Vis. Comput. 29 (2013) 983–1009.

[20] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M.J. Black, Towards understanding action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3192–3199.

[21] Y. Kong, Y. Fu, Human action recognition and prediction: A survey, Int. J. Comput. Vis. 130 (5) (2022) 1366–1401.

[22] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, Z. Li, Towards unified surgical skill assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9522–9531.

[23] Z. Wang, Z. Gao, L. Wang, Z. Li, G. Wu, Boundary-aware cascade networks for temporal action segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, Springer, 2020, pp. 34–51.

[24] Y.A. Farha, J. Gall, Ms-tcn: Multi-stage temporal convolutional network for action segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3575–3584.

[25] Y. Cheng, Q. Fan, S. Pankanti, A. Choudhary, Temporal sequence modeling for video event detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2227–2234.

[26] D.-A. Huang, L. Fei-Fei, J.C. Niebles, Connectionist temporal modeling for weakly supervised action labeling, in: Computer Vision–ECCV 2016: 14th European Conference, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 137–153.

[27] B. Singh, T.K. Marks, M. Jones, O. Tuzel, M. Shao, A multi-stream bi-directional recurrent neural network for fine-grained action detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1961–1970.

[28] D. Wang, D. Hu, X. Li, D. Dou, Temporal relational modeling with self-supervision for action segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 2729–2737.

[29] Y. Huang, Y. Sugano, Y. Sato, Improving action segmentation via graph-based temporal reasoning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14024–14034.

[30] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[34] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[35] S. Karaman, L. Seidenari, A. Del Bimbo, Fast saliency based pooling of fisher encoded dense trajectories, in: ECCV THUMOS Workshop, Vol. 1, 2014, p. 5.

[36] H. Kuehne, J. Gall, T. Serre, An end-to-end generative framework for video segmentation and recognition, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–8.

[37] S. Bhattacharya, M.M. Kalayeh, R. Sukthankar, M. Shah, Recognition of complex events: Exploiting temporal dynamics between underlying concepts, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2235–2242.

[38] C. Lea, M.D. Flynn, R. Vidal, A. Reiter, G.D. Hager, Temporal convolutional networks for action segmentation and detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 156–165.

[39] M.-H. Chen, B. Li, Y. Bao, G. AlRegib, Z. Kira, Action segmentation with joint self-supervised temporal domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9454–9463.

[40] N. Behrmann, S.A. Golestaneh, Z. Kolter, J. Gall, M. Noroozi, Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation, in: European Conference on Computer Vision, Springer, 2022, pp. 52–68.

[41] B. van Amsterdam, A. Kadkhodamohammadi, I. Luengo, D. Stoyanov, Aspnet: Action segmentation with shared-private representation of multiple data sources, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2384–2393.

[42] S.-J. Li, Y. AbuFarha, Y. Liu, M.-M. Cheng, J. Gall, Ms-tcn++: Multi-stage temporal convolutional network for action segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2020).

[43] F. Yi, H. Wen, T. Jiang, Asformer: Transformer for action segmentation, 2021, arXiv preprint arXiv:2110.08568.

[44] D. Du, B. Su, Y. Li, Z. Qi, L. Si, Y. Shan, Do we really need temporal convolutions in action segmentation, 2022, p. 13, arXiv preprint arXiv:2205.13425 2.

[45] J. Wang, Z. Wang, S. Zhuang, H. Wang, Cross-enhancement transformer for action segmentation, 2022, arXiv preprint arXiv:2205.09445.

[46] Y. Ishikawa, S. Kasai, Y. Aoki, H. Kataoka, Alleviating over-segmentation errors by detecting action boundaries, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2322–2331.

[47] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 4904–4916.

[48] M. Wang, J. Xing, Y. Liu, Actionclip: A new paradigm for video action recognition, 2021, arXiv preprint arXiv:2109.08472.

[49] M. Li, L. Chen, Y. Duan, Z. Hu, J. Feng, J. Zhou, J. Lu, Bridge-prompt: Towards ordinal action understanding in instructional videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19880–19889.

[50] S. Stein, S.J. McKenna, Combining embedded accelerometers with computer vision for recognizing food preparation activities, in: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2013, pp. 729–738.

[51] H. Kuehne, A. Arslan, T. Serre, The language of actions: Recovering the syntax and semantics of goal-directed human activities, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 780–787.

[52] A. Fathi, X. Ren, J.M. Rehg, Learning to recognize objects in egocentric activities, in: CVPR 2011, IEEE, 2011, pp. 3281–3288.

[53] C. Lea, A. Reiter, R. Vidal, G.D. Hager, Segmental spatiotemporal cnns for fine-grained action segmentation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, Springer, 2016, pp. 36–52.

[54] C. Lea, R. Vidal, G.D. Hager, Learning convolutional action primitives for fine-grained action recognition, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 1642–1649.

[55] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017, arXiv preprint arXiv:1711.05101.

[56] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (11) (2008).

[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[58] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12009–12019.