Temporal Micro-Action Localization for Videofluoroscopic Swallowing Study

Xianghui Ruan, Meng Dai, Zhuokun Chen, Zeng You, Yaowen Zhang, Yuanqing Li, *Fellow, IEEE*, Zulin Dou, and Mingkui Tan, *Senior Member, IEEE*

***Abstract*—Videofluoroscopic swallowing study (VFSS) visualizes the swallowing movement by using X-ray fluoroscopy, which is the most widely used method for dysphagia examination. To better facilitate swallowing assessment, the temporal parameter is one of the most important indicators. However, most information of that acquire is hand-crafted and elaborated, which is time-consuming and difficult to ensure objectivity and accuracy. In this article, we propose to formulate this task as a temporal action localization task and solve it using deep neural networks. However, the action of VFSS has the following characteristics such as small motion targets, small action amplitudes, large sample variances, short duration, and variations in duration. Furthermore, all existing methods often rely on daily behaviors, which makes locating and recognizing micro-actions more challenging. To address the above issues, we first collect and annotate the VFSS micro-action dataset, which includes 847 VFSS data from 71 subjects, due to the lack of benchmarks. We then introduce a coarse-to-fine mechanism to handle the short and repeated nature of micro-actions, which can significantly enhancing micro-action localization accuracy. Moreover, we propose a Variable-Size Window Generator method, which improves the model's characterization performance and addresses the issue of different action timings, leading to further improvements in localization accuracy. The results of our experiments demonstrate the superiority of our method, with significantly improved performance (46.10% vs. 37.70%).**

***Index Terms*—Temporal parameters, videofluoroscopic swallowing, micro-action, temporal action localization.**

Manuscript received 21 April 2023; revised 3 August 2023; accepted 31 August 2023. Date of publication 8 September 2023; date of current version 6 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62072190 and STI 2030-Major Projects 2022ZD0208900, in part by Program for Guangdong Introducing Innovative and Enterpreneurial Teams under Grant 2017ZT07X183. *(Xianghui Ruan, Meng Dai, and Zhuokun Chen contributed equally to this work.) (Corresponding authors: Yuanqing Li; Zulin Dou; Mingkui Tan.)*

Xianghui Ruan is with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China, also with the Guizhou Minzu University, Guiyang 550025, China, and also with Pazhou Laboratory, Guangzhou 510320, China (e-mail: 202010107625@mail.scut.edu.cn).

Meng Dai, Yaowen Zhang, and Zulin Dou are with the Department of Rehabilitation Medicine, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China (e-mail: daim@mail3.sysu.edu.cn; zhangyw86@mail.sysu.edu.cn; douzulin@mail.sysu.edu.cn).

Zhuokun Chen and Mingkui Tan are with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: 202221045817@mail.scut.edu.cn; mingkuitan@scut.edu.cn).

Zeng You is with the School of Future Technology, South China University of Technology, Guangzhou 510006, China, and also with Peng Cheng Laboratory, Shenzhen 510320, China (e-mail: sezengyou@mail.scut.edu.cn).

Yuanqing Li is with the College of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China, and also with Pazhou Laboratory, Guangzhou 510320, China (e-mail: auyqli@scut.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2023.3313255, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2023.3313255

I. Introduction

SWALLOWING is a complicated physiological reflex process that is at present generally divided into three phases: the oral phase, the pharyngeal phase, and the esophageal phase [1]. There may be subtle disturbances in the function of any one of these phases that can eventually cause dysphagia. Many diseases, such as Parkinson's disease [2], esophageal cancer [3], and stroke [4], have swallowing disorders as early symptoms. As the most widely used and well-researched assessment technique for swallowing disorders is the videofluoroscopic swallowing study (VFSS) and regarded as the gold standard [5]. Specifically, it is imaging of the swallowing movements of the oropharynx, larynx, and esophagus, performed under X-ray fluoroscopy, which is then played back slowly frame-by-frame to analyze and detect abnormalities in swallowing capabilities.

In the VFSS, the temporal parameters are one of the most important indicators used to assess swallowing disorders [6]. These parameters are generally divided into two types: Durations(the length of time for a distinct physiological swallow event to occur) [7] and Intervals (the length of time between two gestures in the swallow sequence) [8], [9]. In clinical applications, they are mainly obtained by manual frame-by-frame observation of the videofluoroscopy by physicians, which is a time-consuming and laborious process. In addition, it is difficult to ensure objectivity and accuracy, since manual observation is restricted by the doctors' experience and professional ability. Therefore, there is a strong need to model the micro-acton in videofluoroscopic swallowing, so that the temporal parameters can be automatically acquired. Inspired by the rapid development of deep learning-based temporal action localization (TAL) methods, we propose the use of the TAL paradigm to model the micro-actions in videofluoroscopic swallowing, enabling the automatic acquisition of the temporal parameters. The TAL task is primarily concerned with locating the times at which the action instances start and end and classifying the action instances within long videos [10]. Furthermore, all existing TAL methods

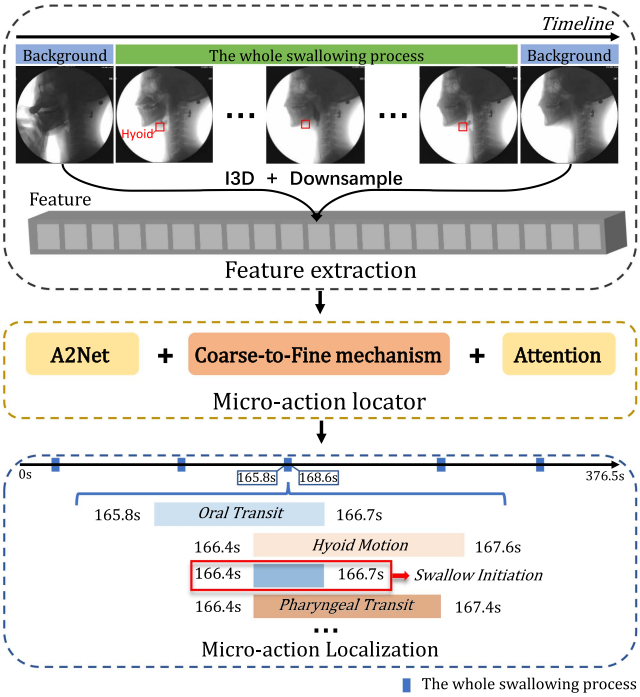

Fig. 1. Pipeline of the proposed temporal micro-action localization for videofluoroscopic swallowing study.

often focus on recognizing everyday behaviors or sports events in general videos [11]. In this article, we attempt to apply the TAL paradigm to the identification of the micro-actions during swallowing in the videofluoroscopic swallowing context. However, it is very challenging to reach this goal for the following reasons.

First, the VFSS micro-action localization is a fine-grained TAL task. The organ movements are small in magnitude, and there is some overlap between them, which makes it difficult to distinguish between different micro-actions. As shown in Fig. 1 the hyoid motion has a small amplitude, and the pharyngeal transit and hyoid motion overlap to a certain extent. To distinguish these micro-actions, we need to focus on more video information for fine-grained localization. Additionally, to obtain features that are more sensitive to local changes, we adopt the method local–global temporal encoder (LGTE) to enhance the feature frame features using the neighboring features of each feature frame. Notably, however, for individual micro-actions with different durations, using only a fixed window size would limit the ability of the local temporal encoder in the attention mechanism to locally model different micro-actions. In this article, we propose the method Variable-Size Window Generator (VSWG), which automatically learns the size of the feature frame region from the features using the self-attentive mechanism feature enhancement method. Second, unlike general actions, swallowing actions are short and occupy only small segment of the entire video; in addition, each micro-action has a different duration. For example, the entire video has a total duration exceeding 370 seconds. However, it only contains 5 clips of the whole swallowing process, each lasting only a few seconds. Among these clips, the oral transit action has a duration of 0.9 seconds, while the swallow initiation is only 0.3 seconds. It is difficult for traditional one-stage methods that directly extract information from the entire video. Specifically, we accordingly propose a coarse-to-fine action localization mechanism that consists of two stages: the first stage localizes the entire swallowing action, and the second stage localizes the timing of the seven stages of micro-actions within that swallowing action. In summary, our main contributions are as follows:

- To the best of our knowledge, this is the first time that a deep learning TAL paradigm has been used to model micro-actions in VFSS for automatic temporal parameter capture.
- We propose a coarse-to-fine swallowing micro-action localization mechanism, which can effectively improve the micro-action localization effect. Furthermore, based on this localization mechanism, we propose a Variable-Size Window Generator method to enhance the features. Experimental results show that our methods achieve 46.1% in terms of mAP, which is 8.4% higher than the baseline when tIou=0.5.
- We create the first dataset for micro-action TAL of VFSS, by performing data cleaning, desensitization, and data enhancement on a total of 847 VFSS from 71 subjects.

The remainder of this article is organized as follows. In Section II, we discuss the related literature, including that pertaining to VFSS methods and deep learning-based TAL. Section III provides a detailed overview of our methods and the VFSS dataset. In Section IV, we present the experiments and results performed with the micro-actions TAL methods. Finally, we conduct a short discussion and then conclude this article.

II. Related Work

In this section, we summarize some of the progress made in VFSS methods and temporal action localization (TAL) methods based on deep learning.

A. Videofluoroscopic Swallowing Study (VFSS)

VFSS methods can be divided into qualitative analysis [12] and quantitative analysis [13]. Qualitative analysis refers to the clinician's observation and assessment of the patient's swallowing function according to his or her own experience. This is relatively easy to perform, efficient, and therefore widely used in current clinical practice; however, it relies heavily on the clinician's experience and imaging quality and is highly subjective.

Quantitative analysis requires the measurement and quantification of temporal and kinematic parameters involved in the swallowing process, making it more comprehensive and precise. However, to obtain accurate parameters, quantitative analysis requires a large number of manual markers, and accordingly a complex process of measuring and quantifying parameters that is very difficult to apply in clinical settings. To reduce the effort associated with quantitative analysis, Kellen et al. [14] manually label the hyoid region and use Sobel edge detection to track the hyoid to calibrate the the motion of hyoid. Hoaasin et al. [15] propose a semi-automatic method to automatically

identify the region of interest before identifying the hyoid, restricting the image processing to the region of interest, then tracking the hyoid and quantifying its motion. Kim et al. [16] propose a deep learning-based segmentation network to determine the motion trajectory of the hyoid bone. Lee et al. [13] used a 3D convolutional neural network to classify swallowing video clips and automatically detect the pharyngeal phase, Lee et al. [17] develop a deep learning-based VFSS program that can automatically detect both the presence of penetration or aspiration swallowing disorders from swallowing videos among patients with dysphagia. However, the above works primarily focused on analyzing specific actions or examining the safety of swallowing, similar to the Action Localization (AL) method, which mainly considers the spatial relationships between objects in video frames to predict action categories and spatial positions. In contrast, our method models the extraction of VFSS temporal parameters as a task of temporal micro-action localization, using deep learning methods to automatically extract these time-related parameters. Similar to the TAL method, we primarily focus on the temporal relationships between frames in the video to predict action categories and temporal boundary information.

B. Deep Learning-Based Temporal Action Localization

Deep learning has many areas of application in health informatics, including bioinformatics, medical imaging, pervasive sensing, medical informatics, and public health [18]. The TAL paradigm we adopted to model the micro-motion localization method in VFFS is one such application in the field of medical imaging. At present, there are two main categories of deep learning-based TAL methods: skeleton-based and video content-based methods. In skeleton-based TAL methods, action is described by the change in the position of human joints between the frames of a video as a result of the changes in the skeleton sequences [19]. Song et al. [20] utilize the information from the human skeleton to extract the most discriminating features of human actions, thereby improving the action recognition performance. Elkholy et al. [21] use body skeleton information for detecting and identifying abnormal human behavior. Skeleton data is not affected by background illumination, and it is robust; however, performance depends on how well it is extracted. Video content-based methods can be further divided into anchor-based and anchor-free localization methods. The anchor-based localization methods are top-down models [10], [22], [23], [24], [25], which can be subdivided into one-stage and two-stage methods. The one-stage methods predict each temporal position and classification simultaneously by anchoring frames (SSAD [22], GTAN [23], GCN [26], MGG [24]). The two-stage method R-C3D [25], GCM [27] first proposes action candidate frames and then classifies the candidate frame regions. Inspired by the Faster R-CNN Framework [10], since the anchor frames are predetermined, the performance of the anchor-based temporal localization method is more dependent on the prior knowledge of the action distribution. For their part, the anchor-free localization methods employ a bottom-up mechanism that detects the actions in the video by predicting an action score for each frame [28], [29], [30], [31]. Each action is a point and then regresses the distance from the point to the start and end boundaries. The first purely anchor-free temporal localization method, AFSD [28], in addition to methods such as SSN [29], BSN [30], and BMN [31], introduces a confidence map on top of BSN [30] that is capable of densely predicting the boundary matching situation. Although the action probability of each frame can provide a useful guide for pinpointing the action boundaries, the incorrect prediction of a particular frame can have a substantial adverse impact on the overall action localization. In micro-action temporal localization, Mi et al. [32] propose a dual-stream convolutional network based on hierarchical pyramids to localize and recognize micro-actions. Bandini et al. [33] release a new dataset for facial motion analysis in individuals with neurological disorders and analysis of face alignment bias. Philipp et al. [34] constructed dietary gesture datasets and applied state-of-the-art action recognition methods for gesture detection. Notably, however, although many temporal localization methods have been proposed in recent years, most of them are implemented based on generic datasets; in contrast, relatively little research has been conducted on the modeling of micro-actions as well as localization recognition.

III. Proposed Method

In this article, the automatic acquisition of the chronological parameters is achieved by the positioning of the temporal micro-action. In the micro-action localization task, the I3D model [35] is used in the video feature extraction phase to extract the optical flow features along with the RGB features of VFSS. The captured features are combined in the channel dimension and fed into the A2Net method [36]. The starting and ending times of each action are first directly predicted using an end-to-end approach, after which a coarse-to-fine mechanism is used to optimize the experimental results, Moreover, a method of self-attention mechanism with variable window size is proposed to enhance the video features and thereby optimize the localization of micro-actions.

A. Problem Definition

Given an untrimmed swallowing video $V = \{v_i\}_{i=1}^{l}$ with l frames, acquisition of the temporal parameters in quantitative analysis is achieved by locating the starting and ending times of swallowing micro-actions in the VFSS. For simplicity, we denote $B(j) = \{t_{s_j}, t_{e_j}, c_j\}$ as all action instances in this video, where t_{s_j} and t_{e_j} are the starting time and ending time of the j-th micro-action temporal parameters, and c_j is the j-th category label.

The analysis of temporal parameters in quantitative VFSS has been mainly divided into durations and intervals [8], [9].

Duration time can also be defined as the period of time that an anatomical structure will be in action within a single swallow, and is used to examine the functional effect that this particular structure will have. Interval time in swallowing refers to the time between the actions of two organ structures and is typically used to reflect the temporal sequence of the swallowing process. The temporal parameters of VFSS are shown in the Temporal Parameters column of Table I.

TABLE I
THE DETAILS OF SWALLOWING MICRO-ACTION CATEGORY

Micro-action category	Start Points	End Points	Temporal Parameters	Average Time (s)	Sample Variance
Oral Transit [7]	Start of oral phase	End of oral phase	Durations	0.865	0.252
Soft Palate Elevation	Start of soft palate elevation	End of soft palate elevation	Intervals	1.334	0.210
Hyoid Motion [8]	Start of hyoid motion	Return of hyoid to original	Durations	1.674	0.588
UES Opening [8]	Start of UES open	UES closed	Durations	0.786	0.071
Swallow Initiation	Start of hyoid motion	End of oral phase	Intervals	0.334	0.106
Pharyngeal Transit [7]	Start of hyoid motion	UES closed	Intervals	1.332	0.228
Laryngeal Vestibule Closure [9]	Closure of laryngeal vestibule	Reopening of laryngeal vestibule	Durations	0.769	0.104

B. VFSS Temporal Micro-Action Localization Dataset

1) Data Acquisition and Labeling: All the videos of fluoroscopic swallowing were accomplished by the dynamic digital radiography machine for swallowing video/image acquisition at the Third Affiliated Hospital of Sun Yat-Sen University. which was digitally recorded as videos (format.AVI) by using the VFSS Acquisition and Analysis system (Longest Inc., Guangzhou, China) at 30 frames/sec. All videos were independently assessed by two experienced clinicians. The total duration of the videos is 508 minutes and 21 seconds collected from 71 subjects. Among them are 25 males and 46 females. We used a sliding window approach for data enhancement to complete the dataset. In more detail, for the one-stage approach, 32 seconds and 64 seconds were used as the sliding window length, and one-fourth of its length was used as the sliding step to intercept video clips from the original video file. All video clips were required to contain at least one complete swallowing process; otherwise, the video clips were invalid. A total of 847 video clips were obtained after the above operation, and the RGB input data needed for feature extraction could also be obtained.

For each swallow, the key time of the start and end points required for the temporal parameters were manually checked frame by frame. Through these time points, the temporal coordinates of the seven types of micro-actions, as well as the complete swallowing process in the original video could be located [8]. The specific relevant correspondences are presented in Table I. At the same time, it can be seen that there is a certain overlap between the micro-actions. We performed a statistical analysis of the average duration and sample variance for each micro-action in the dataset in Table I. It can be seen that the average duration of all micro-action is relatively short, with the longest being the hyoid motion at only 1.674 s and the shortest being the swallow initiation at only 0.334 s. The variances of the different micro-actions also vary, with the largest variance being 0.588 and the smallest only 0.071. This indicates that the duration of the micro-actions in the videofluoroscopic swallowing study is short and inconsistent. Considering the characteristics of the videofluoroscopic swallowing micro-actions, it is difficult to localize the final results directly from the whole video.

2) Data Pre-Processing: The production of datasets mainly involves the production of one-stage model datasets and the training of coarse-to-fine model datasets. The one-stage approach requires predicting the start time, end times, and category labels of each micro-action directly from the whole video. Moreover, training of the one-stage model needs to be performed on the dataset of a relatively long video. For this part, the coarse-to-fine strategy first predicts the complete swallowing process from the video and then predicts the micro-actions from the swallowing process; thus, the model generally needs to be trained on a shorter video. The collected dataset is divided into training data, validation data, and test data in a ratio of 4:1:1. We performed dataset partitioning on a per-subject basis, ensuring that each person's swallowing processes appeared in only one dataset, thereby avoiding data leakage issues. For the coarse-to-fine localization mechanism dataset, the sliding window length is set to 4 seconds and the sliding window step is set to 3 frames. The sliding window sampling process of the one-stage model is also repeated to eventually obtain 9788 videos of 4 seconds in length after the cut, and the annotation information is completed in the same ways as the one-stage dataset processing process for the coarse-to-fine localization mechanism.

3) Feature Representation: Most current feature representation methods for action recognition use RGB frames and optical flow to capture appearance and motion information through two-stream networks [35], [37]. In this work, we used two-stream [35] for the extraction of the input video features for encoding. For a given video, RGB features are extracted for every video clip, and a matrix of size $T\times$W $\times$ H $\times$ 3 is obtained for each video. For each 8 frames of the matrix, the feature frames are taken once in 3 steps to obtain the input data of $b \times 8\times$W $\times$ H $\times$ 3, where b is the number of feature frames. The time dimension T of the matrix is uniformly adjusted to 256 by linear interpolation, and the batchsize is set to 32; the result is then input into the I3D model [35] to obtain a 256 $\times$ 1024 feature vector. Both the RGB features and optical flow features have the same dimensionality. These features are then incorporated into the dataset file, finalizing the training model dataset.

C. Coarse-to-Fine Localization Mechanism

Due to the short duration of micro-actions to be localized in the video perspective swallowing temporal action localization (TAL) task (less than one second on average), along with the overlapping area between micro-actions, it is difficult to localize the temporal information of micro-actions from the complete video. To solve the above problem, we adopt a coarse-to-fine localization mechanism to improve the effect, as shown in Fig. 2. First, the whole swallowing process is localized from the video, after which the swallowing candidate frames are filtered with confidence thresholds; the features are then re-extracted from the corresponding segments of the swallowing candidate frames, and the extracted features are input separately into the single-class localizer to predict the temporal information of the corresponding micro-action categories. Finally, the predicted candidate frame positions are summed with the offset of the

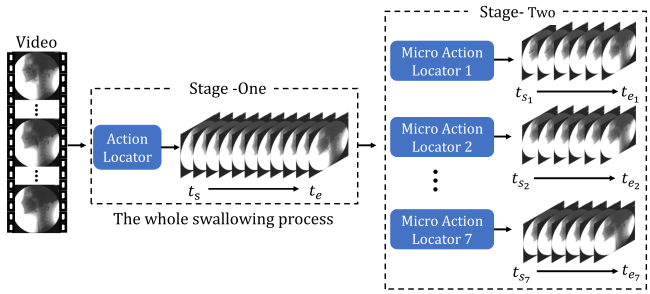

Fig. 2. Proposed acquisition of temporal parameters for videofluoroscopic swallowing using a coarse-to-fine temporal action localization mechanism.

localized video itself to obtain the results of the method. The detailed process is as follows: First, the TV-L1 method [38] is used to extract optical flow frames, and the RGB data and optical flow data are separately input into the I3D model to obtain the dual- stream features of the input video. Similarly to the feature pyramid approach in object detection, to improve the ability of the temporal action localization model to locate actions at different temporal scales, dual stream features are first input into a basic convolutional layer for dimensionality reduction. The obtained features are then input into the LGTE module [39] optimized by VSWG to obtain enhanced features. These enhanced features are further processed by a hierarchical feature module consisting of six layers with different kernel sizes, resulting in six scale-specific feature maps.

We use both an anchor-free module and an anchor-based module to predict potential candidate boxes for each scale-specific feature map. Specifically, we divide the video into six equal-length parts (128, 64, 32, 16, 8, 4), considering the start and end positions of each part as the predefined temporal boundaries for the anchor-free module. The center point position and temporal duration of each part are considered as the anchor center point position and temporal duration for the anchor-based module. For the anchor-free module, we use the six-scale features to predict the boundary offsets for each scale. For the anchor-based module, we use the six-scale features to predict the center points and temporal duration offsets for each scale. We then merge the predicted candidate boxes from both branches and apply non-maximum suppression to remove redundant candidate boxes, resulting in the final predictions of the one-stage method. In the proposed method, we use A2Net [36] to train a single-class localizer for each micro- action category in a training set of shorter video duration, and moreover adopt a stage-two localization approach based on the stage-one action detectors.

D. Self-Attention Method With Variable Window Size

Due to the short duration of micro-actions in the VFSS with localization tasks, it is necessary to acquire features that are more sensitive to changes between nearby frames. To enhance the video features so that they are more representative of the changes between neighboring frames and the differences with the global data of a particular frame, we apply a variable self-attention mechanism [40], combining LGTE [39] and VSWG, to enhance video features for micro-actions. This mechanism is built upon a coarse-to-fine localization approach, aiming to improve the effectiveness of localization.

1) Self-Attention Method for Feature Enhancement: The features are divided into eight parts from the feature dimension, four parts for local temporal encoder (LTE) processing and four parts for global temporal encoder (GTE) processing, and the neighboring window size of LTE is set to 41.

First, the region around the feature frame is intercepted using a specific size window, and the feature frame and the feature vector are linearly mapped with a learnable matrix and then self-attention is applied to enhance the features within its surrounding region, resulting in locally enhanced features. Furthermore, the feature vector of the feature frame is linearly mapped with another learnable matrix and then subjected to generate globally enhanced features. At the same time, the feature frames are input into the VSWG module to obtain the variable window weights W_i^v, and W_i^v is multiplied with the local features obtained from LTE, after which the LTE results are restricted to the predicted window range by means of a mask. The enhanced local features are concatenated together with the global features and multiplied with a learnable matrix; subsequently, the obtained results are passed through a feed-forward neural network to increase the nonlinear parameters to further enhance the feature expression. Finally, the enhanced features are obtained. More details as shown in Fig. 3.

$$f_i^l = \text{Attention}\left(\gamma^l(f_i), \rho^l(f_w), \varphi^l(f_w)\right) \tag{1}$$

where f_i is the feature of the i-th feature frame in the video feature, f_w is the feature frame corresponding to the neighboring window segment, f_i^l is the feature of the i -th feature frame following enhancement with neighboring features, and γ^l, ρ^l, and φ^l are linear mapping functions that can be learned.

$$f_i^g = \text{Attention}\left(\gamma^g(f_i), \rho^g(f_v), \varphi^g(f_v)\right) \tag{2}$$

The GTE calculation process is similar to LET, where f_v is the feature of the whole video, f_i^g is the feature of the i-th feature frame after enhancement with global features, and γ^g, ρ^g, and φ^g are linear mapping functions that can be learned.

2) Variable-Size Window Generator: Considering the different durations of individual micro-actions, using only a fixed-size window size will limit the ability of the local temporal encoder (LTE) to locally model different micro-actions. In this article, we accordingly propose the Variable-Size Window Generator method (VSWG) to for optimization. This method utilizes feature frame context information to dynamically learn the length of the region to be attended by the feature frames from the features. Specifically, for each feature frame f_i of a video feature, a module consisting of an average pooling layer, leakyrelu, fully connected layer, and sigmoid is used to predict the appropriate one-way window size threshold for the current feature th_i, which can be calculated as follows:

$$th_i = \lfloor (W+1)/2 \rfloor \cdot \text{Sigmoid}(\text{Linear}(\text{LeakyReLU}(\text{AveragePool}(f_i)))) \tag{3}$$

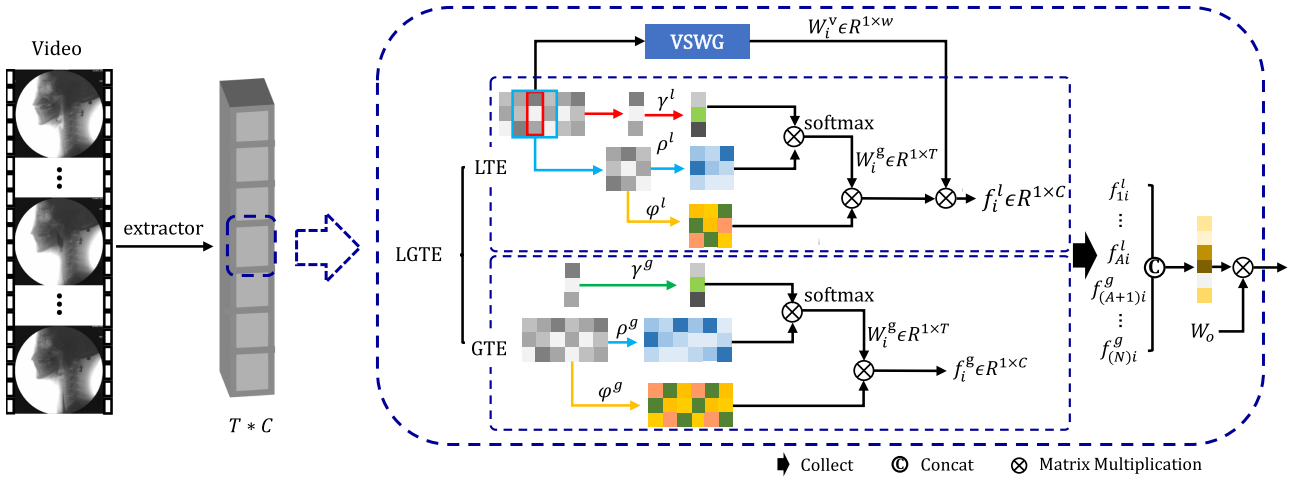

Fig. 3. Detailed structure of the Self-attention method with variable window size enhancing features. The acquisition of the feature frames is divided into two groups: one input to the global temporal encoder, and the other to the VSWG method that is integrated into the local temporal encoder.

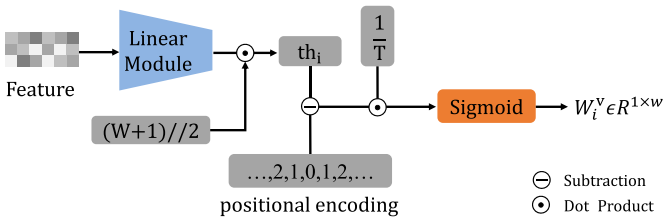

Fig. 4. Detailed operating of Variable-Size Window Generator method.

where W is the preset maximum window size and th_i takes the value range $[0, \lfloor (W+1)/2 \rfloor]$.

$$p = [\lfloor (W-1)/2 \rfloor, \ldots, 1, 0, 1, \ldots, \lfloor (W-1)/2 \rfloor] \quad (4)$$

$$e_i = th_i - p \quad (5)$$

$$W_i^v = \text{Sigmoid}\,(e_i/T) \quad (6)$$

The value of e_i is used to indicate whether each position within the window range is within the predicted window size threshold. The sigmoid function is then used to threshold e_i and obtain an approximate one-hot encoded weight sequence W_i^v. Here T is a control parameter of the sigmoid function, and the value of W_i^v is closer to one-hot encoding when T is smaller; more details of its operation are presented in Fig. 4.

IV. Experiments

In this section, we compare the baseline and coarse-to-fine micro-action localization mechanism and present the experimental results on a VFSS localization task.

A. Implementation Details

1) One-Stage Micro-Action Localization: We used the I3D model [35] as the network for feature extraction. The obtained optical flow, as well as the RGB features, were stitched together in the feature dimension to form 2048-dimensional feature values, which were then input into the model for localization. Considering that there is a large overlap rate among micro-actions in the swallowing video, and moreover that the multi-classification cannot completely distinguish between and localize micro-actions with a high overlap rate, we use the A2Net model [36] for TAL on the swallowing video. The single-classification and multi-classification methods are then used to localize the swallowing videos. Finally, the effects of the two localization methods are compared.

2) Coarse-to-Fine Micro-Action Localization: To improve the micro-action localization performance, we conducted a coarse-to-fine experiment based on the baseline method. In more detail, we select the baseline model with the best localization performance as the prediction model in the first stage of the coarse-to-fine mechanism, then use the produced video clips to train the localization model of micro-action classes required for the second stage.

The whole swallowing process is first localized using the overall swallowing action classification of the first stage. Subsequently, the action proposals with a confidence level greater than 0.5 in the localization results of the first stage are selected, after which the time dimension of the obtained action proposals centroids is extended outward to 4 seconds to obtain a new interval containing complete swallowing candidate boxes as input to the model. These data then need to be uniformly extended to 256 in the time dimension by the linear difference method, after which they are input into the I3D feature extraction model [35], and the obtained dual-stream features are input into the single-class classification of each micro-action. Finally, the micro-action localization results are obtained.

B. Evaluation Metric

We use the evaluation metric mean average precision (mAP) from ActivityNet for temporal action localization (TAL), where

TABLE II
PERFORMANCE COMPARISONS WITH MULTI CLASSIFICATION AND SINGLE CLASSIFICATION VIDEO TEMPORAL ACTION LOCALIZATION IN ONE-STAGE METHODS, MEASURED BY MAP(%)

Method	Micro-action	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Multi-classification	Oral Transit	**42.3**	**41.9**	34.1	26.4	13.1	7.0	2.3
	Soft Palate Elevation	72.1	66.8	63.0	48.8	30.0	9.6	3.2
	Hyoid Motion	**60.0**	**57.5**	**53.3**	**39.6**	**25.6**	**13.4**	**4.5**
	UES Opening	41.2	38.8	26.0	13.7	4.0	0.9	0.7
	Swallow Initiation	12.0	9.0	3.9	0.9	0.3	0.1	0.1
	Pharyngeal Transit	**76.8**	**76.8**	**75.4**	66.8	48.4	24.6	7.8
	Laryngeal Vestibule Closure	**72.1**	**69.8**	**63.9**	**55.8**	**35.5**	18.6	6.2
	Average	53.7	51.5	45.7	36.0	22.4	10.6	3.5
Single-classification	Oral Transit	35.9	35.9	**35.9**	**30.8**	**20.7**	**9.0**	**2.9**
	Soft Palate Elevation	**79.4**	**79.4**	**79.4**	**78.2**	**63.5**	**34.6**	**13.4**
	Hyoid Motion	36.3	36.3	36.3	31.5	10.7	3.5	1.7
	UES Opening	**74.6**	**74.4**	**74.4**	**73.1**	**62.0**	**44.7**	**17.5**
	Swallow Initiation	**33.4**	**29.2**	**27.3**	**22.1**	**12.0**	**4.8**	**0.3**
	Pharyngeal Transit	73.4	73.4	73.4	**72.1**	**61.3**	**38.5**	**13.3**
	Laryngeal Vestibule Closure	53.4	53.4	53.2	49.1	33.7	**19.4**	**7.9**
	Average	**55.2**	**54.6**	**54.3**	**51.0**	**37.7**	**22.1**	**8.1**

The bold numbers indicate the best results.

AP is the evaluation metric for each action class and mAP is the average result of APs from multiple classes. In the evaluation strategy, the confidence scores of the localized candidate frames are ranked, and the top 100 are taken to calculate the final mAP. All the anchor frames are sorted from highest to lowest in order of confidence size. From this, we can calculate the sequence of anchor box recall and accuracy, which are computed as follows:

$$recall = \frac{TN}{TP + FN} \tag{7}$$

$$precision = \frac{TP}{TP + FP} \tag{8}$$

where TP is the current number of positive samples, FP is the number of negative samples, and FN is the number of undetected positive samples. The precision-recall curve can be obtained by taking the recall rate as the horizontal axis and the precision as the vertical axis. We calculate the area of the curve and the coordinate axis to get the final value of AP.

C. Experimental Results Analysis

We first compare the effect of the localization method with single-classification and multi-classification methods. Table II summarizes the comparison performances on our dataset. In terms of the specific micro-action localization effect, although some micro-actions are better localized in the multi-method than single under threshold 0.1:0.2, the single classification model localization effect is mostly better than multi-classification under threshold 0.3:0.7.

The overall micro-actions localization mAP of the single-classification method under threshold 0.5, the pharyngeal transit, soft palate elevation, and UES opening micro-actions could reach more than 60. In contrast, hyoid motion and swallow initiation had the worst effect, at only 10.7 and 12.0; even under threshold 0.1, they could only perform 36.3 and 33.4. More details are shown in Table II. Since the videofluoroscopic swallowing task can be performed offline and does not require high inference speed, we choose the single-classification method as the baseline for optimization.

To analyze the reasons for the differences in the localization performance between different micro-actions, we can see in Tables I and II that among the types of actions with positioning effects greater than 60.0, pharyngeal transit and soft palate elevation were above 1 s, and the mean duration of UES opening was short, but its variance was only 0.071; moreover, the mean duration of swallow initiation among the actions with the worst positioning effects was only 0.334 s, and the mean duration of hyoid motion was 1.674 s, although its variance was the largest. This indicates that the difference in localization effect among the different micro-actions is caused by the difference in the mean duration and variance, which means the short and unstable duration that leads to the poor localization effect of some micro-actions. Considering that it is difficult to locate the micro-action intervals directly from the whole video due to their relatively short duration of micro-action, the optimization of the model localization effect using a coarse-to-fine localization mechanism is again considered. As can be seen from Table III, the coarse-to-fine method achieves a significant approach that improves the localization effect of each micro-action compared to the one-stage method from Table II. The coarse-to-fine improves the performance from 33.7 to 43.3 measured by average mAP under an IoU threshold of 0.5 and is most significantly improved when the IoU threshold is less than 0.5. Notably, when the IoU threshold was 0.1, the coarse-to-fine method was able to achieve 70.5, an improvement of about 15 relative to the baseline method. These results indicate that the coarse-to-fine method is more effective than the baseline method in "finding" and "fixating" the action. Under threshold 0.1:0.7, using LGTE alone does not optimize the coarse-to-fine method, and there is a significant decrease in performance at multiple IoU thresholds. More detailed information can be found in the Table III. This occurs because different types of micro-actions have different durations, and there are large duration fluctuations for the same micro-action; therefore, using a fixed attention window of fixed length cannot effectively obtain information from the video to enhance the features. Moreover, the localization effect of the model is further improved after we incorporate LGTE & VSWG into the coarse-to-fine approach: mAP is improved by 2.9 under threshold 0.4 and 2.8 under threshold 0.5.

TABLE III
PERFORMANCE COMPARISONS WITH THE METHODS, MEASURED BY MAP(%)

Method	Micro-action	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Coarse-to-Fine	Oral Transit	**75.7**	69.2	54.9	35.8	16.8	10.6	**7.4**
	Soft Palate Elevation	69.4	69.4	69.4	69.4	67.3	43.1	**15.2**
	Hyoid Motion	73.8	73.8	65.0	**36.4**	16.3	7.8	1.1
	UES Opening	**71.9**	**71.9**	**71.9**	**71.9**	70.8	56.5	36.4
	Swallow Initiation	**60.6**	36.9	**20.6**	8.8	3.1	1.1	5.0
	Pharyngeal Transit	72.4	72.4	72.4	**72.4**	63.0	31.6	15.0
	Laryngeal Vestibule Closure	69.8	69.8	69.8	69.8	65.9	54.5	**32.8**
	Average	70.5	66.2	60.6	52.1	43.3	29.3	15.5
Coarse-to-Fine & LGTE	Oral Transit	73.7	65.9	47.2	28.0	18.1	8.4	2.8
	Soft Palate Elevation	**73.4**	**73.4**	**73.4**	**73.4**	**72.4**	**58.5**	12.5
	Hyoid Motion	74.4	74.4	52.6	35.4	16.9	**8.8**	**9.0**
	UES Opening	71.8	71.8	71.8	71.8	69.4	**57.6**	**42.5**
	Swallow Initiation	52.7	32.8	15.6	8.1	1.9	1.4	6.0
	Pharyngeal Transit	**73.3**	**73.3**	**73.3**	71.6	54.5	25.9	15.6
	Laryngeal Vestibule Closure	71.2	71.2	71.2	71.2	**71.2**	50.7	26.6
	Average	70.1	66.1	57.8	51.3	43.5	30.2	14.5
Ours	Oral Transit	75.1	**72.2**	**59.8**	38.7	**23.0**	**11.4**	7.1
	Soft Palate Elevation	71.7	71.7	71.7	71.2	66.3	42.9	10.1
	Hyoid Motion	**75.4**	**75.4**	**71.8**	50.9	**27.9**	7.8	1.1
	UES Opening	71.7	71.7	71.7	71.7	**71.2**	57.5	42.1
	Swallow Initiation	58.7	37.7	19.3	**8.8**	**3.8**	**1.5**	**8.0**
	Pharyngeal Transit	71.9	71.9	71.9	71.9	**65.0**	**43.7**	**19.8**
	Laryngeal Vestibule Closure	**71.7**	**71.7**	**71.7**	**71.7**	65.6	**56.6**	29.4
	Average	**70.9**	**67.5**	**62.5**	**55.0**	**46.1**	**31.6**	**15.8**

The bold numbers indicate the best results.

TABLE IV
EFFECT OF DIFFERENT WINDOWSIZES ON THE LOCALIZATION PERFORMANCE OF MICRO-ACTIONS, MEASURED BY AVERAGE MAP (%)

WindowSize	0.3	0.5	0.7	Average
5	57.5	42.3	12.5	37.4
9	58.7	43.3	14.5	38.8
11	58.8	44.9	**16.5**	40.1
21	59.2	44.7	13.2	39.0
31	57.1	42.4	14.7	38.1
41	**62.5**	**46.1**	15.8	**41.5**
45	60.4	42.1	14.5	39.0
51	60.3	44.9	15.7	40.3

The bold values indicate the best results.

This indicates that the enhancement of video features using the self-attention approach is effective. Moreover, from the data in Table III, it can be seen that the hyoid motion (with the largest variance and the most mAP) is improved by 14.5 under threshold 0.4 and 11.6 under threshold 0.5. Accordingly, the coarse-to-fine architecture achieves a significant increase at all thresholds and is also the most significant among several types of micro-actions, which indicates that the enhanced features of our method have a better enhancement effect for the more variable motions. Several other action types with smaller variances also exhibit significant improvement.

D. What WindowSize Should Be Set

In Table IV, we conducted experiments to explore the impact of different WindowSize settings ranging from 5 to 51. We observed that minimal WindowSize values led to an inadequate collection of local details, limiting the model's effectiveness in capturing essential features. Conversely, when the WindowSize was set to an extremely large value, it introduced excessive global noise, negatively impacting overall performance. To strike a balance between capturing local details and avoiding excessive noise, we carefully evaluated the results for various WindowSize settings. Among all the tested values, we found that setting the WindowSize to 41 yielded the best performance for the seven micro-actions in videofluoroscopic swallow studies. Therefore, we ultimately set the WindowSize to 41 in our experiments.

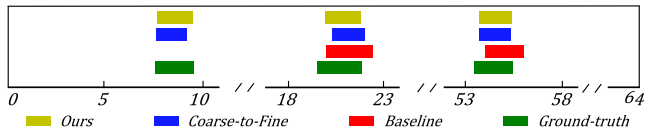

Fig. 5. Visualization of the effect of three methods of localization (a).

E. Qualitative Results

Given the significant improvements brought about by our method, we next attempt to provide a more visual demonstration of its effective role. We randomly select a video from the test dataset and visualize results for the baseline method, the coarse-to-fine localization architecture, and the effect of micro-action localization after optimization with ours in Fig. 5. For the baseline method, localization of micro-action from the entire 64 s duration video is very difficult, and the baseline method prediction results have a true value without any prediction with overlap to match it in the first micro-action positioning in the video.

In the coarse-to-fine method prediction results, the ground truths in the videos all have some overlap with the predicted values, and the overlap range with the ground truth is greater than the effect of the baseline method; this reflects that the coarse-to-fine architecture is better at finding micro-actions compared with the baseline method. In addition, it can be seen from the figure that for those micro-actions that have been localized, the

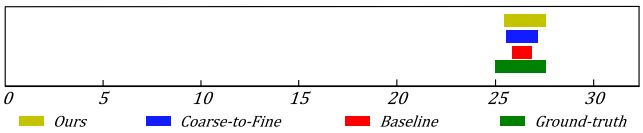

Fig. 6. Visualization of the effect of three methods of localization (b).

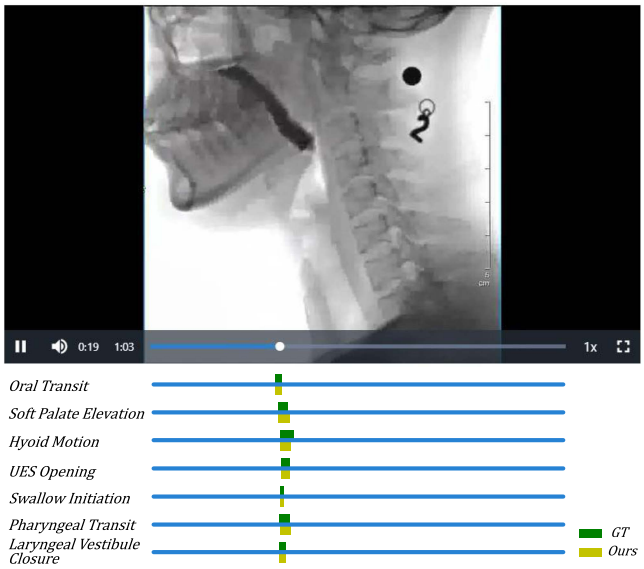

Fig. 7. Visualization of the effect of our method on micro-actions.

model is more accurate in localizing the boundaries following the optimization of our feature enhancement.

We also visualize the effect from an entire video of 32 s in duration. As shown in Fig. 6, overlap with ground truth can be found in the prediction results of both the baseline and coarse-to-fine localization methods. The visualization results further show that the coarse-to-fine prediction results are closer to the ground truth values than the baseline, which intuitively reflects that the coarse-to-fine method is more accurate in regressing the micro-actions boundaries than the baseline method. In addition, we also visualize the localization effect of our method on micro-actions, as shown in Fig. 7, our method can more accurately localize each swallowing micro-action. For example, our method for localization of the swallow initiation is very close to Ground truth(GT) in a short duration, and the model is more accurate for TAL on micro-actions. The acquisition of time parameters of each VFSS is more accurate.

V. Discussion

In this article, we propose a method TAL for automatically obtaining time parameters in VFSS. Swallowing disorders are a common medical issue that can severely impact a patient's quality of life. Accurately diagnosing these disorders is crucial for effective treatment, and VFSS is a widely used diagnostic tool. However, manually analyzing VFSS videos is time-consuming and requires specialized expertise. The TAL method is a significant contribution to the field, as it can automatically extract temporal information from VFSS videos, reducing the workload of medical professionals and increasing the accuracy of diagnoses. Meanwhile, we also introduce a new dataset containing videos of swallowing micro-actions, along with their corresponding clinical analyses. This dataset is the first of its kind and is expected to have a significant impact on the field. It can facilitate the development of new methods for acquiring temporal parameters in VFSS and the creation of intelligent complementary medical systems. It can also serve as a benchmark dataset for future research.

To improve the effectiveness of the I3D model in feature extraction on the swallowing micro-action dataset, we fine-tune the model and perform a series of exploratory experiments on VFSS. These experiments provided valuable insights into the micro-action TAL accuracy. As shown in Tables II and III, these results support the use of the coarse-to-fine & LGTE& VSWG method.

The proposed method achieves high accuracy in extracting temporal information from VFSS videos, providing a promising approach for the automated diagnosis of swallowing disorders. In the future, we plan to further enrich the dataset by annotating key points for swallowing organs and incorporating key point sequence information to guide micro-action feature extraction. These improvements are expected to enhance micro-action localization and improve the accuracy of temporal parameter acquisition, enabling more precise diagnoses of swallowing disorders. Overall, this article presents a valuable contribution to the field of swallowing disorder diagnosis and treatment, and we believe it will have a significant impact on future research in this area.

VI. Conclusion

In this article, we developed the first VFSS micro-action localization dataset. Moreover, the existing feature extraction and temporal localization model is applied to obtain the temporal parameters of the swallowing video by fine-tuning the temporal localization of the seven micro-actions in a videofluoroscopic swallowing video. We experimentally demonstrated that using the coarse-to-fine method outperformed the one-stage method by a large margin (43.3% vs. 37.7%).

For different individuals between whom there are large fluctuations in the duration of VFSS micro-actions, we propose a Variable-Size Window Generator method based on the existing self-attention mechanism to enhance video features and thereby enhance the model's representation ability. Finally, the proposed method achieves a result of mAP@0.5=46.1.

Therefore, new VFSS micro-action localization datasets being made available may enable automated assessments of swallowing disorders to become more accurate.

References

[1] K. Matsuo and J. B. Palmer, "Anatomy and physiology of feeding and swallowing: Normal and abnormal," *Phys. Med. Rehabil. Clin. North Amer.*, vol. 19, no. 4, pp. 691–707, 2008.

[2] C. Pflug et al., "Critical dysphagia is common in parkinson disease and occurs even in early stages: A prospective cohort study," *Dysphagia*, vol. 33, pp. 41–50, 2017.

[3] R. E. Martin, P. Letsos, D. H. Taves, R. Inculet, H. X. Johnston, and H. G. Preiksaitis, "Oropharyngeal dysphagia in esophageal cancer before and after transhiatal esophagectomy," *Dysphagia*, vol. 16, pp. 23–31, 2014.
[4] S. M. Hughes, "Management of dysphagia in stroke patients," *Nurs. Older People*, vol. 23, no. 3, 2011, Art. no. 21.
[5] Y. J. Na, J. S. Jang, K. H. Lee, Y. J. Yoon, M. S. Chung, and S. H. Han, "Thyroid cartilage loci and hyoid bone analysis using a video fluoroscopic swallowing study (VFSS)," *Medicine*, vol. 98, no. 30, 2019, Art. no. e16349.
[6] K. A. Kendall, S. McKenzie, R. J. Leonard, M. I. Gonçalves, and A. Walker, "Timing of events in normal swallowing: A videofluoroscopic study," *Dysphagia*, vol. 15, no. 2, pp. 74–83, 2000.
[7] R. O. Dantas, R. de Aguiar Cassiani, C. M. Dos Santos, G. C. Gonzaga, L. M. T. Alves, and S. C. Mazin, "Effect of gender on swallow event duration assessed by videofluoroscopy," *Dysphagia*, vol. 24, no. 3, pp. 280–284, 2009.
[8] S. M. Molfenter and C. M. Steele, "Temporal variability in the deglutition literature," *Dysphagia*, vol. 27, no. 2, pp. 162–177, 2012.
[9] T. Park, Y. Kim, D.-H. Ko, and G. McCullough, "Initiation and duration of laryngeal closure during the pharyngeal swallow in post-stroke patients," *Dysphagia*, vol. 25, no. 3, pp. 177–182, 2010.
[10] Y. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1130–1139.
[11] P. Chen et al., "RSPNet: Relative speed perception for unsupervised video representation learning," in *Proc. AAAI Conf. Innov. Appl. Artif. Intell.*, 2021, pp. 1045–1053.
[12] Y.-C. Chang et al., "Dysphagia in patients with nasopharyngeal cancer after radiation therapy: A videofluoroscopic swallowing study," *Dysphagia*, vol. 18, no. 2, pp. 135–143, 2003.
[13] J. T. Lee, E. Park, and T.-D. Jung, "Automatic detection of the pharyngeal phase in raw videos for the videofluoroscopic swallowing study using efficient data collection and 3D convolutional networks," *Sensors*, vol. 19, no. 18, 2019, Art. no. 3873.
[14] P. M. Kellen, D. L. Becker, J. M. Reinhardt, and D. J. Van Daele, "Computer-assisted assessment of hyoid bone motion from videofluoroscopic swallow studies," *Dysphagia*, vol. 25, no. 4, pp. 298–306, 2010.
[15] I. Hossain, A. Roberts-South, M. Jog, and M. R. El-Sakka, "Semi-automatic assessment of hyoid bone motion in digital videofluoroscopic images," *Comput. Methods Biomech. Biomed. Eng. Imag. Vis.*, vol. 2, no. 1, pp. 25–37, 2014.
[16] H.-I. Kim, Y. Kim, B. Kim, D. Y. Shin, S. J. Lee, and S.-I. Choi, "Hyoid bone tracking in a videofluoroscopic swallowing study using a deep-learning-based segmentation network," *Diagnostics*, vol. 11, no. 7, 2021, Art. no. 1147.
[17] S. Lee, S. Choi, J. Ko, and H. Kim, "Deep learning based application for videofluoroscopic swallowing study(VFSS): A pilot study," *J. Neurological Sci.*, vol. 405, pp. 78–79, 2019.
[18] D. Raví et al., "Deep learning for health informatics," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.
[19] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1110–1118.
[20] S. Song, J. Liu, Y. Li, and Z. Guo, "Modality compensation network: Cross-modal adaptation for action recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 3957–3969, 2020.
[21] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba, "Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 1, pp. 280–291, Jan. 2020.
[22] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 988–996.
[23] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 344–353.
[24] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S.-F. Chang, "Multi-granularity generator for temporal action proposal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3604–3613.
[25] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5783–5792.
[26] R. Zeng et al., "Graph convolutional networks for temporal action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7093–7102.
[27] R. Zeng et al., "Graph convolutional module for temporal action localization in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6209–6223, Oct. 2022.
[28] C. Lin et al., "Learning salient boundary feature for anchor-free temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3320–3329.
[29] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2914–2923.
[30] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[31] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3889–3898.
[32] Y. Mi and S. Wang, "Recognizing micro actions in videos: Learning motion details via segment-level temporal pyramid," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 1036–1041.
[33] A. Bandini et al., "A new dataset for facial motion analysis in individuals with neurological disorders," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 1111–1119, Apr. 2021.
[34] P. V. Rouast and M. T. P. Adam, "Learning deep representations for video-based intake gesture detection," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 6, pp. 1727–1737, Jun. 2020.
[35] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.
[36] L. Yang, H. Peng, D. Zhang, J. Fu, and J. Han, "Revisiting anchor mechanisms for temporal action localization," *IEEE Trans. Image Process.*, vol. 29, pp. 8535–8548, 2020.
[37] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[38] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L 1 optical flow," in *Proc. 29th DAGM Symp. Pattern Recognit.*, 2007, pp. 214–223.
[39] Z. Qing et al., "Temporal context aggregation network for temporal action proposal refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 485–494.
[40] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.