Preprints.org

Article

Not peer-reviewed version

End-to-End Swallowing Event Localization via Blue-Channel-to-Depth Substitution in RGB-D: GNConvNeXt-Modified AdaTAD with KAN-Chebyshev Decoder

Derek Ka-Hei Lai , Zi-An Zhao , Andy Yiu-Chau Tam , Jing Li , Jason Zhi-Shen Zhang , Duo Wai-Chi Wong * , James Chung-Wai Cheung *

Posted Date: 3 October 2025

doi: 10.20944/preprints202510.0294.v1

Keywords: temporal action detection; computer vision; dysphagia screening; swallowing detection; Kolmogorov-Arnold Networks

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

End-to-End Swallowing Event Localization via Blue-Channel-to-Depth Substitution in RGB-D: GRNConvNeXt-Modified AdaTAD with KAN-Chebyshev Decoder

Derek Ka-Hei Lai [1], Zi-An Zhao [1], Andy Yiu-Chau Tam [1], Jing Li [1], Jason Zhi-Shen Zhang [2], Duo Wai-Chi Wong [1,*] and James Chung-Wai Cheung [1,3,*]

[1] Department of Biomedical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong

[2] Department of Mathematics, School of Science, Hong Kong University of Science and Technology, Hong Kong

[3] Research Institute of Smart Ageing, The Hong Kong Polytechnic University, Hong Kong

* Correspondence: duo.wong@polyu.edu.hk (D.W.-C.W.); james.chungwai.cheung@polyu.edu.hk (J.C.-W.C.)

Abstract

Swallowing is a complex biomechanical process, and its impairment (dysphagia) poses major health risks for older adults. Current diagnostic methods such as videofluoroscopic swallowing (VFSS) and fiberoptic endoscopic evaluation of swallowing (FEES) are effective but invasive, resource-intensive, and unsuitable for continuous monitoring. This study proposes a novel end-to-end RGB–D framework for automated swallowing event localization in continuous video streams. The framework enhances the AdaTAD backbone through three key innovations: (i) an RGD input strategy that substitutes the noisy blue channel with depth information to capture subtle neck movements, (ii) a GRNConvNeXtAdapter that incorporates Global Response Normalization for efficient temporal feature adaptation, and (iii) a Kolmogorov–Arnold Network (KAN) decoder that leverages Chebyshev polynomials for non-linear temporal modeling. (iv) boundary regression head and sinusoidal patch embedding to further optimize model precision. Evaluation on a proprietary swallowing dataset comprising 641 clips and 3,153 annotated events demonstrated the effectiveness of the proposed framework. We analysed and compared modification strategy across design of adapter, decoder, input channel combination, regression method, and patch embedding technique. The optimized configuration (VideoMAE + GRNConvNeXtAdapter + KAN + RGD + boundary regression + sinusoidal embedding) achieved an average mAP of 83.25%, significantly surpassing the baseline I3D + RGB + MLP model (61.55%). Ablation studies further confirmed that each architectural component contributed incrementally to the overall improvement. These results establish the feasibility of accurate, non-invasive, and automated swallowing event localization using depth-augmented video. The proposed framework paves the way for practical dysphagia screening and long-term monitoring in clinical and home-care environments.

Keywords: temporal action detection; computer vision; dysphagia screening; swallowing detection; Kolmogorov-Arnold networks

1. Introduction

Dysphagia, characterized by difficulty in swallowing, is a significant geriatric syndrome affecting more than half of older adults with dementia and is associated with a 13-fold increase in mortality risk [1,2]. This condition impairs the natural swallowing process, which occurs approximately 200 – 1000 times daily in healthy individuals [3], and can lead to severe complications

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

including aspiration pneumonia, which is the third leading cause of injury deaths in older people [4]. The prevalence of dysphagia ranges from 25% in adults to 60% in residential care facilities [5,6], imposing substantial economic burdens with median hospitalization charges exceeding $30,000 for aspiration pneumonia cases [7]. Current diagnostic approaches using videofluoroscopic swallowing (VFSS) and fiberoptic endoscopic evaluation of swallowing (FEES) are considered gold standards but present significant limitations including invasiveness, patient discomfort, radiation exposure, high costs, and requirements for specialized personnel [8]. These constraints make these assessment unsuitable for routine screening or continuous monitoring.

As dysphagia represents a gradual aging process that deteriorates progressively with cognitive decline and neurological disorders [9], continuous monitoring becomes essential to identify high-risk stages for timely intervention. Wearable sensor-based screening systems accompanied with machine learning or deep learning enable continuous monitoring using sensors such as inertial measurement units (IMUs) with accelerometers and gyroscopes, electromyography (EMG) electrodes, acoustic sensors, microphones, nasal airflow sensors, and strain sensors [10–13]. Although these systems show promise as alternatives to traditional methods, they face significant challenges, including poor signal quality, susceptibility to motion artifacts, and patient compliance issues, particularly among older adults with dementia [10–13]. In contrast, computer vision represents an emerging trend that addresses these limitations without requiring direct physical contact.

Computer vision approaches have emerged predominantly through conventional RGB camera systems for non-contact dysphagia monitoring [11]. Sakai, et al. [14] developed a machine learning-based screening test using image recognition from iPad cameras to assess sarcopenia dysphagia. Similarly, Yamamoto, et al. [15] employed a compact 3D camera to detect lip motion patterns and quantify swallowing dynamics during bolus flow in elderly participants. However, the performance of RGB-based systems has been moderate [14]. This moderate performance can be attributed to the inherent limitation that swallowing processes involving hyoid bone movement, laryngeal elevation, and subtle soft tissue deformation in the neck region, are often too subtle for visual observation through conventional RGB cameras alone. RGB-D (depth) cameras offer significant advantages by providing three-dimensional (3D) spatial information that can capture subtle surface deformations and movement patterns invisible to conventional cameras. Lai, et al. [16] successfully demonstrated this approach using RGB-D cameras to create a comprehensive dataset of swallowing activities and achieved an accuracy (F1-score) of 92% using Transformer X3D.

Despite its promising potential to address contact-based sensor limitations, computer vision for dysphagia screening faces significant technical challenges that hinder its transition to real-world monitoring. Current computer vision approaches, whether utilizing RGB or RGB-D modalities, are predominantly constrained to pre-designed experimental protocols, such as the Comprehensive Assessment Protocol for Swallowing (CAPS) [17], that focus on controlled swallowing and non-swallowing (such as reading and dry-swallowing) tasks rather than naturalistic eating behaviors. These studies typically require manual temporal segmentation and windowing of video sequences, with researchers manually clipping specific swallowing events from longer recordings for analysis. The absence of automated event detection in continuous video streams severely limits the potential for ubiquitous screening scenarios, such as capturing and analyzing entire meal-taking processes without human oversight.

Recent advances in computer vision have increasingly focused on temporal action localization (TAL) models, which aim to precisely identify and temporally segment specific actions within continuous video streams. TAL represents a fundamental paradigm shift from traditional frame-based classification approaches toward comprehensive temporal understanding, enabling systems to automatically detect when actions occur and determine their precise temporal boundaries without manual intervention [18]. The evolution of TAL approaches has progressed through several distinct paradigms, from early two-stage methods to sophisticated end-to-end. One-step approaches aim to directly predict actions at the frame level without generating proposals, offering simplicity but often struggling with long-range temporal dependencies. A notable example is the Convolutional-De-

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

Convolutional (CDC) network, introduced by Shou, et al. [19], which places CDC filters on top of 3D ConvNets that are effective for abstracting action semantics but reduce temporal length. Two-step approaches involve generating temporal proposals and then classifying them. The paradigm is exemplified by the Structural Segment Network that models the temporal structure of each action instance via a structured pyramid and introduces a decomposed discriminative model with an action classifier and completeness detector [20].

An end-to-end architecture for TAL is typically realized by integrating a feature extractor with a localization head. A prominent example of this paradigm is ActionFormer, a state-of-the-art single-stage detector [21]. It leverages a Transformer-based encoder to process multiscale feature representations, employing local self-attention to efficiently model temporal context, followed by a lightweight decoder that classifies each temporal moment and regresses the starting and ending boundaries of the action [21]. Conventionally, ActionFormer operates on features pre-extracted by an offline backbone, most notably I3D (Inflated 3D ConvNet) [22]. I3D adapts 2D convolutional kernels for 3D spatiotemporal data by "inflating" them, enabling it to effectively learn motion patterns from video and produce powerful, generic action features [22]. To create a more cohesive end-to-end system, recent work like AdaTAD (Adapter Tuning for Temporal Action Detection) has explored merging ActionFormer with adaptable backbones such as VideoMAE (Video Masked Autoencoder) [23,24]. VideoMAE, a self-supervised model pre-trained via a masked autoencoding strategy, excels at learning robust and instance-specific representations [24]. The key innovation of AdaTAD is its adaptive framework, which allows the VideoMAE backbone to be efficiently fine-tuned alongside the ActionFormer detector for the specific target dataset. However, this ambition of full end-to-end fine-tuning introduces significant practical limitations. The computational demands and memory requirements of jointly training these sophisticated architectures are substantial, and attempting to update all parameters of a large pre-trained backbone like VideoMAE risks the catastrophic forgetting of its valuable, generalized representations.

This challenge has driven the development and adoption of Parameter-Efficient Fine-Tuning (PEFT) approaches as a solution [25]. In fact, the AdaTAD framework is a prime example of this strategy in action [23]. Instead of full fine-tuning, PEFT methods enable the targeted adaptation of pre-trained models by updating only a small subset of parameters or by inserting lightweight, trainable modules (e.g., adapters) into the frozen backbone. This approach allows the model to specialize in the downstream task while preserving the integrity of majority of its pre-trained knowledge.

Most existing TAL models are designed and pre-trained exclusively on three-channel RGB inputs, making it challenging to integrate modalities such as depth without significant architectural modifications or retraining, especially in specialized domains like swallowing event localization To address this limitation, our study proposes a novel framework that fine-tunes the model by thoroughly examining five major components: the learnable adapter, the decoder, input combinations, the regression method, and patch embedding techniques. We validate our approach using a custom-built RGB-D video dataset of swallowing events, which we have meticulously collected and annotated for this purpose.

2. Materials and Methods

2.1. Overview

This study introduces a novel, end-to-end framework for localizing swallowing events in continuous RGB-D video streams. Our approach is built upon the AdaTAD architecture, which couples a VideoMAE feature extractor with an ActionFormer-based detector. We systematically enhance this baseline by exploring modifications across five key architectural components:

1. Temporal Feature Adapter: We conduct a comparative analysis of five Parameter-Efficient Fine-Tuning (PEFT) adapters: the original BottleneckAdapter, InvertedConvNeXtAdapter,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

GRNConvNeXtAdapter, Adapter+, and Compacter, to optimize temporal feature learning while the VideoMAE backbone remains frozen.
2. Decoder Head: We replace the standard Multilayer Perceptron (MLP) decoder in the detection head with a Kolmogorov-Arnold Network (KAN) that utilizes Chebyshev polynomials, hypothesizing it can better model the non-linear dynamics of swallowing.
3. Input Modality: To leverage depth information, we evaluate a channel substitution strategy (RGD, RDB, DGB) and compare its performance against standard RGB and traditional early/late fusion RGB-D methods.
4. Regression Method: We compare two strategies for boundary prediction: centerness-based regression and direct boundary regression.
5. Patch Embedding: We assess the impact of different positional encoding techniques, specifically comparing the standard sinusoidal positional encoding with Rotary Positional Encoding (RoPE).

The methodology unfolds in two main phases. First, we perform an initial ablation study on the public THUMOS14 dataset to benchmark the performance of different adapter and decoder combinations. Second, we transition to our proprietary swallowing dataset, where we evaluate the complete set of modifications to identify the optimal configuration for domain-specific swallowing event detection. Model performance is quantified using mean Average Precision (mAP) at various temporal Intersection over Union (tIoU) thresholds, with the final proposed model's efficacy validated against baseline models.

2.2 Data Acquisition and Labelling

Data was collected from 136 older adults at day centers and care homes. Half of the participants (n = 68) were clinically diagnosed with dysphagia and followed a dysphagia diet at International Dysphagia Diet Standardization Initiative (IDDSI) level 4 or below. The other 68 participants had no dysphagia. Individuals with a history of neck surgery, tracheostomy, or feeding tube use were excluded. Additionally, participants with cognitive impairments that hindered their ability to understand, respond to, or comply with study requirements were deemed ineligible. The mean age of participants was 85 ± 7.4 years.

During the data acquisition phase, movements of the lower face and neck, including the lips, mandible, and throat, were recorded during swallowing and non-swallowing tasks using an RGB-D camera (Intel RealSense D435i, Intel Corp., Santa Clara, USA). As illustrated in Figure 1, participants were seated in a neutral position with an eye-level marker on a table serving as a fixed reference point to ensure consistent posture. The camera, set to 30 frames per second with a resolution of 480 × 848 pixels, was positioned on a table approximately 35 cm from the target anatomical regions and angled at 45° relative to the horizontal plane to optimize visibility of the neck and mandible. All data were stored on a connected computer. We adapted the Comprehensive Assessment Protocol for Swallowing (CAPS) [17], but restricted the testing boluses to IDDSI levels 0 to 4 to reduce choking risks associated with more solid textures, in accordance with the recommendations from our occupational therapist. Participants performed tasks that included swallowing foods of varying textures or non-swallowing activities, such as coughing or speaking, with each video capturing the same action repeated five times consecutively.

Following data collection, an occupational therapist conducted a thorough manual review of the footage to identify and isolate clips with clear depictions of swallowing and non-swallowing actions. The therapist manually marked the timeframes for swallowing and non-swallowing events in each video based on observations of the RGB and depth (D) video data, clipping the footage to focus on these events. Videos were excluded if the camera view was obstructed, if involuntary participant movements occurred during feeding, or if the precise temporal onset of the swallowing event could not be determined. This curation process resulted in a final dataset of 641 complete video clips for subsequent analysis.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

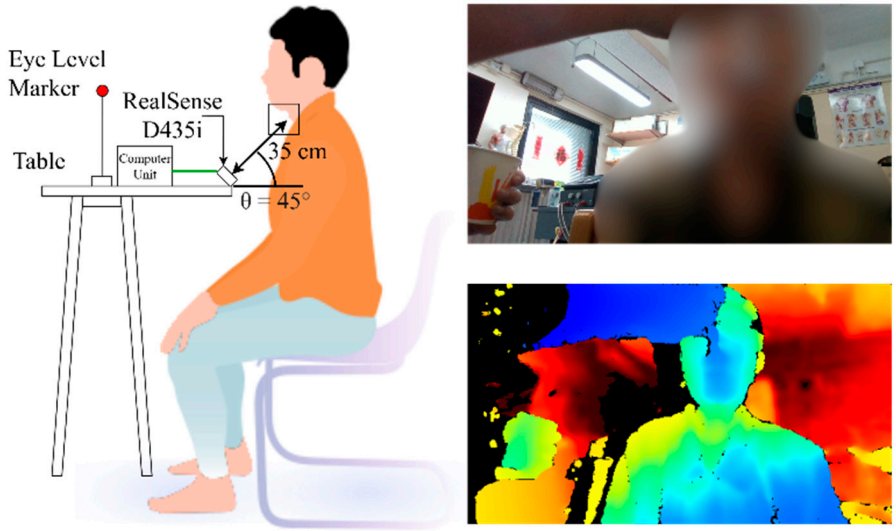

Figure 1. Experimental Setup for Data Collection. The participant was seated in a neutral position, with an RGB-D camera placed on a table approximately 35 cm from the target anatomical regions (lips, mandible, and throat) and angled at 45° relative to the horizontal plane. The left side of the figure shows an illustrative screenshot displaying both the RGB and depth (D) channel outputs.

2.3. Data Processing

RGB data were utilized directly without additional processing. For depth data, we applied the processing pipeline recommended by the Intel RealSense SDK to optimize data quality. Initially, depth data were transformed into the disparity domain using a Depth-to-Disparity transform, which facilitated subsequent filtering. To reduce spatial noise, an edge-preserving spatial filter was applied in the disparity domain. Temporal consistency was enhanced, and noise further minimized, through a temporal filter that leveraged information from previous frames in the disparity domain. Finally, the processed data were converted back to the depth domain using a Disparity-to-Depth transform, preparing them for subsequent analysis. Depth data were clipped at 1.0 m to suppress irrelevant background disparity and focus on the near-field neck region, then rescaled to 8-bit to match the dynamic range of RGB inputs. This choice standardized channel ranges while preserving salient geometric features. Empirically, we observed negligible impact of alternative clipping thresholds (e.g., 0.8 or 1.2 m) or higher quantization, indicating robustness of the normalization strategy.

2.4. Baseline Model Architecture

We developed an end-to-end TAL architecture through our modifications to the AdaTAD framework, as baseline model illustrated in Figure 2. For feature extraction, we leverage VideoMAE (Vision Transformer ViT) as the foundational backbone and pretrained on the Kinetic-400 dataset [26].

To enhance training efficiency and reduce memory usage, we implemented a parameter-efficient fine-tuning (PEFT) approach that involves freezing the encoder of VideoMAE, thereby preserving its powerful pre-trained representations, and strategically tuning a feature adapter [27]. This adapter is crucial for capturing the dynamic temporal characteristics inherent in swallowing events.

To optimize the model for our data, we explored modifications on different parts, which include exploring the best adapter, comparing detector head using MLP with KAN and finding the best input channel combination. Further, after finding the best adapter, we fine-tuned the adapter by searching for the optimal depth, scaling factor, with addition of residual connections.

We adopted AdaTAD for its ability to capture long-range dependencies via local self-attention and hierarchical temporal pyramids, which scale efficiently compared to recurrent–convolutional hybrids. While recurrent–convolutional systems remain viable [28], their sequential recurrence limits

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

parallelism and increases training time. Our adoption of AdaTAD harnesses transformer efficiency and enables parameter-efficient fine-tuning, achieving superior performance across numerous public dataset benchmarks, it demonstrates superior results in many public dataset benchmarking tools. Nevertheless, we will retain ActionFormer with I3D feature extraction as the baseline for model comparison.

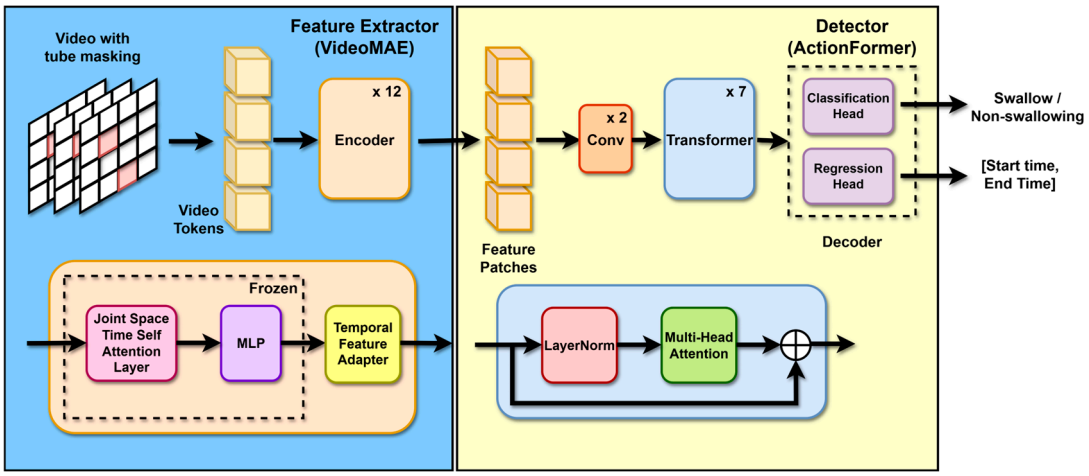

Figure 2. Architecture of The Baseline AdaTAD. The model comprises a feature extractor utilizing the VideoMAE-B backbone with a parameter-efficient fine-tuning (PEFT) encoder, followed by different temporal feature adapters. The detector, adapted from the original ActionFormer, includes a classification head and a regression head, implemented as a multilayer perceptron (MLP) via 1D convolution.

2.5. Adapter Exploration

We compare five recently proposed adapter designs.

BottleneckAdapter. This module follows the compress–process–expand paradigm originally proposed for ResNet architectures and used in the AdaTAD baseline as the temporal-informative adapter [23]. It begins with a downscaling convolution that reduces the channel dimension to one quarter, followed by a SiLU (Sigmoid Linear Unit) activation (Figure 3a). A depthwise 1D convolution with kernel size 3 then captures local temporal patterns. Channel-wise interactions are introduced via a subsequent pointwise (1×1) 1D convolution. The output of these operations is added to the downscaled features through a residual connection. An up-projection layer restores the original channel dimension, and a second residual connection adds the result to the original input, enabling efficient temporal modeling and stable optimization.

The InvertedConvNeXtAdapter, illustrated in Figure 3b, is inspired by the ConvNeXt block design [29]. It begins with a down-projection layer that reduces the feature dimensionality, followed by a depth-wise 1D convolution with a kernel size of 7, which efficiently captures long-range temporal dependencies across channels. Layer normalization is applied to stabilize training. Next, a patch-wise convolution expands the feature dimensionality fourfold, enhancing representational capacity. A GELU activation introduces non-linearity, and a second patch-wise convolution projects the features back to their original dimension. Finally, an up-projection layer restores the full representation, and a residual connection adds the adapter's input to its output, improving gradient flow and enabling more effective learning.

GRNConvNeXtAdapter. Inspired by ConvNeXt V2 [30], this variant integrates a Global Response Normalization (GRN) layer after the patch-wise convolutions. GRN adaptively normalizes features based on their global response, improving information flow, generalization, and the modeling of long-range dependencies while maintaining stable training. The remainder of the block follows an inverted bottleneck structure, including an initial down-projection, depthwise convolution, GELU activation, and up-projection, with a residual connection to preserve input information.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

Adapter+ [31]. Following the Houlsby-style adapter, Adapter+ removes internal normalization and employs channel-wise scaling to finely modulate the adapter's contribution with negligible overhead.

Compacter [32]. For the final variant, we implement Compacter, which replaces the down- and up-projection layers with two low-rank parameterized hypercomplex multiplication (LPHM) layers. Instead of learning full weight matrices, each LPHM layer constructs its weight matrix as a sum of Kronecker products between shared "slow" weights and adapter-specific "fast" rank-one weights. This reduces parameter complexity from O(kd) in standard adapters to O(k + d). Concretely, A matrices are shared across layers to capture general adaptation knowledge, while B matrices are parameterized in low rank to model layer-specific adaptations.

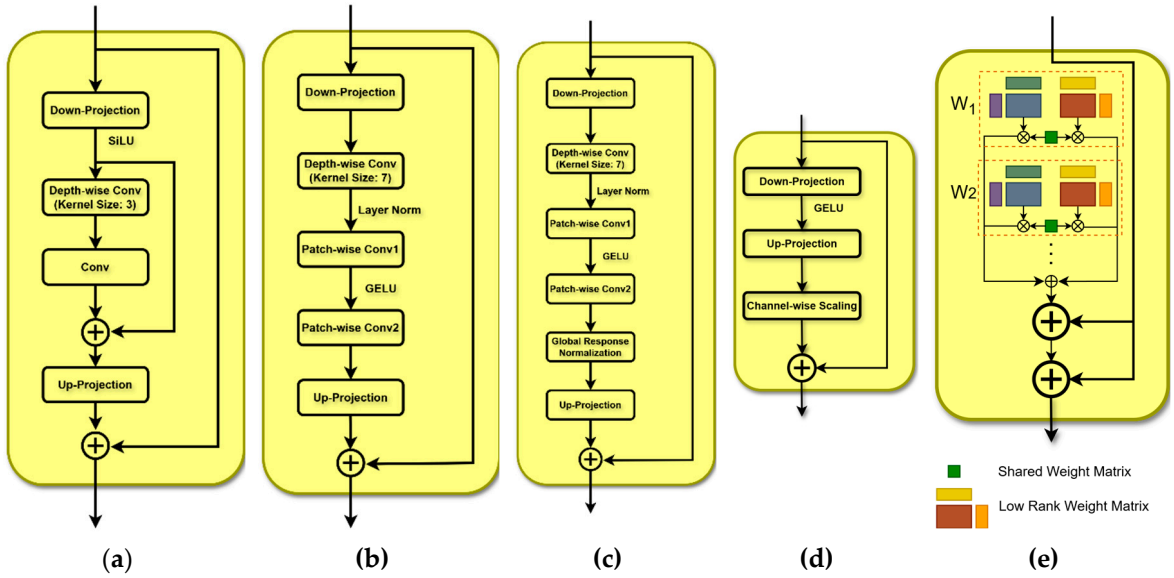

Figure 3. Temporal feature adapter configurations: (a) BottleneckAdapter; (b) InvertedConvNeXtAdapter; (c) GRNConvNeXtAdapter; (d) Adapter+; (e) Compacter.

Following feature extraction, the processed features are fed into the detection component, which is based on the ActionFormer framework (Figure 2). It begins with two convolutional layers to project the input features before passing them to the transformer encoder. Its encoder consists of seven transformer blocks, each employing local attention mechanisms to efficiently capture temporal dependencies. Notably, the last five transformer blocks apply 2× downsampling, progressively reducing temporal resolution to facilitate multi-scale temporal modeling. This architectural design creates a hierarchical feature pyramid across multiple temporal scales. Both the classification and regression heads operate on these multi-level features, enabling precise temporal action localization by capturing and integrating information at different temporal resolutions.

2.6. Decoder Selection

We specifically focus on tuning the decoder within ActionFormer to optimize its performance for event localization. Apart from the baseline multilayer perceptron (MLP) layer, we examine a more recent powerful alternative based on the KAN theorem. The KAN theorem establishes that any multivariate continuous function can be represented as a finite sum of univariate continuous functions, providing a powerful theoretical foundation for modeling complex, high-dimensional relationships [33]. We explored the use of KAN in the decoder based on the hypothesis that it can better capture the inherent spatial and temporal coherence present in video data. This continuity property may enable the KAN-based decoder to exploit underlying structures in the data more effectively, potentially leading to improved model performance in temporal action localization.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

To detail the core transformation, the input is first normalized to the range [−1,1] by applying a hyperbolic tangent activation (1).

$$A = \tanh(X) \tag{1}$$

The main mechanism lies the generation of Chebyshev polynomials as feature expansions [34]. For the normalized input A, the process proceeds as follows: Let k = [0,1,...,D] represent the set of polynomial degrees considered. For each value of k, the Chebyshev polynomial of order k is computed using the identity (2).

$$\theta = \arccos\big(\text{clamp}(A, -1+\epsilon, 1-\epsilon)\big), \qquad P_k = \cos(k \cdot \theta) \tag{2}$$

where ϵ is a small positive value to ensure numerical stability. This step expands the channel dimension of the input from its original value to (D+1) times larger, producing a polynomial feature space P.

A standard convolutional operation with learnable weights W is then applied to P, as shown in (3):

$$Z = \text{Conv}(P; W) \tag{3}$$

Finally, the features are passed through a normalization layer, and optionally a dropout layer, to improve training stability and generalization (4).

$$Y = \text{Dropout}(\text{Norm}(Z)) \tag{4}$$

This module ultimately replaces the traditional 1D convolutional layer of MLP in the model, serving as the core component for classification and regression tasks.

For inference, the model processes input video streams in fixed windows of 768 frames, segmenting each window into 48 non-overlapping temporal chunks, each comprising 16 consecutive frames. These chunks are individually encoded by the backbone network, after which their corresponding feature maps are concatenated along the temporal dimension. To ensure temporal alignment with the original input, the concatenated features are subjected to spatial pooling and then temporally interpolated back to the full 768-frame resolution. The resulting unified feature map is subsequently fed into a streamlined ActionFormer detection head, which outputs precise temporal boundaries for each action instance within the sequence.

For training, the model is optimized using the AdamW algorithm, employing a learning rate schedule that combines an initial linear warm-up phase with subsequent cosine decay. The learning rates are set to 1×10^{-4} for the detection head and 2×10^{-4} for the adapter modules, reflecting the differential adaptation needs of these components.

2.7. Data Input Strategy

The regression head is responsible for precisely localizing the temporal boundaries of a detected swallowing event. To identify the most effective approach for this task, we compared two distinct regression strategies. The first is centerness-based regression, which predicts a single value for each temporal location within a potential action. This "centerness" score quantifies how close a given point in time is to the center of the action instance, effectively down-weighting predictions from locations near the start or end boundaries, which are often less reliable. The second strategy is boundary-based regression, which, for each temporal position, directly predicts two values: the distance to the action's starting boundary (onset) and the distance to its ending boundary (offset). Both regression heads share the same bottleneck architecture as the classification head—comprising three 1D convolutional layers with kernel sizes of 1, 3, and 1—and operate across all levels of the temporal feature pyramid to enable multi-scale boundary prediction. Our experiments systematically evaluated both methods to determine which provided more accurate temporal localization for swallowing events.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

2.8. Data Input Strategy

To utilize depth channel data, we substituted each individual channel with the depth channel, creating RGD, RDB, DGB inputs. Recognizing that the blue channel may offer less salient information in certain imaging scenarios, we hypothesized that incorporating depth would provide more informative spatial cues for localizing swallowing events. The performance of the RGD input was then compared to the standard RGB configuration.

In addition to depth substitution, we implemented early and late fusion baselines. Early fusion applied a convolutional layer to project the four-channel RGBD input into a three-channel representation, preserving compatibility with pre-trained backbones. Late fusion processed RGB and depth streams independently and fused them in feature space before the detection head, mimicking the dimensionality of the RGB backbone output. These baselines provide fair comparisons for evaluating the effectiveness of our substitution strategy.

2.9. Patch Embedding Method

Sinusoidal Positional Encoding: In the context of transformer-based models applied to image processing tasks, such as image classification or object detection, sinusoidal positional encoding serves as a mechanism to incorporate spatial positional information into the model. Transformers, originally designed for sequential data like text, lack an inherent understanding of the two-dimensional spatial arrangement of pixels or patches in images. Sinusoidal positional encoding addresses this by assigning each image patch or pixel to a unique positional vector, generated using fixed, periodic mathematical functions. These vectors create a structured pattern that allows the model to distinguish the relative positions of patches, such as whether one patch is to the left or above another. This is particularly critical in vision transformers, where the model must capture spatial relationships to understand image content effectively. However, the fixed nature of sinusoidal encodings limits their adaptability to diverse image sizes or task-specific spatial patterns, potentially constraining performance in complex visual tasks.

Rotary Positional Encoding (RoPE): Rotary Positional Encoding (RoPE) represents an advanced approach to embedding positional information in transformer models for image processing, offering improved flexibility over traditional sinusoidal encodings. In vision tasks, RoPE integrates positional information by applying rotation transformation to the feature representations of image patches or pixels, effectively encoding their relative spatial relationships. Unlike sinusoidal encodings, which rely on adding separate positional vectors, RoPE modifies the attention mechanism itself by rotating the query and key vectors based on their positions in the image grid. This rotation-based approach preserves the relative distances between patches, enabling the model to better capture spatial dependencies, such as the arrangement of objects in an image. RoPE's ability to generalize across varying image sizes and its computational efficiency make it particularly effective for tasks like image segmentation or scene understanding, where precise spatial awareness is essential, often outperforming sinusoidal encodings in modern vision transformer architectures.

2.10. Model Training

Training used AdamW (lr 1e−4 for detection head, 2e−4 for adapters; weight decay 0.05). The backbone was frozen. A linear warmup over 5 epochs transitioned to cosine decay. Gradient clipping was applied at 1.0. Maximum training was 100 epochs, with convergence reached by epoch 60. Mixed precision was enabled, batch size was 2, and 20 data loader workers were used. EMA tracking and static graph optimization were applied. Experiments ran on Nvidia RTX 4090, with 24GB VRAM. BatchNorm was retained for compatibility with pre-trained weights, despite small batch sizes.

2.11. Model Evaluation

Before adapting our framework to the specific domain of swallowing events, we conducted a comprehensive evaluation and ablation study on the THUMOS14 dataset [35]. This dataset, widely

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

recognized for its action recognition and temporal localization challenges, serves as an ideal benchmark for validating the core components of our modified ActionFormer. The primary objective of this phase is to assess and select the optimal configurations for both the temporal feature adapter (BottleneckAdapter, InvertedConvNeXtAdapter; and GRNConvNeXtAdapter) and the decoder (MLP vs. KAN). I3D (Inflated 3D ConvNet) [22], trained over 35 epochs, served as the baseline model for comparative analysis.

Following the successful benchmarking on THUMOS, we adapted and trained the model on our proprietary RGB-D swallowing dataset. A total of 641 video clips were collected, each with a mean duration of approximately 27 seconds (Table 1). Some clips appeared lengthy because they included both the preparatory and recovery phases of the event. Patients with dysphagia may require extra time to clear any residue, which contributed to the extended duration. These clips contained 3,153 annotated episodes encompassing both swallowing and non-swallowing events. Model evaluation was conducted using a training-testing split of approximately 80% and 20%, respectively. I3D (inflated3D ConvNet) and the original AdaTAD served as the baseline models for comparative analysis

Table 1. Dataset composition and descriptive statistics for the RGB-D swallowing video, stratified by training and testing set.

Data metric	Training	Testing	Total
No. of video clips	512 (79.88%)	129 (20.12%)	641
Total duration (min)	229.6	55.17	284.71
Mean (SD) duration (s)	26.90 (20.92)	25.66 (20.02)	26.65(20.75)
Max duration (s)	150.80	122.22	–
Min duration (s)	4.74	7.04	–
Annotated swallowing events	1091	260	1351
Annotated non-swallowing events	1427	375	1802
Total No. of events	2518 (79.86%)	635 (20.14%)	3153

SD: standard deviation.

Performance was assessed using the mean Average Precision (mAP) metric, computed at multiple temporal Intersection over Union (tIoU) thresholds of 0.1 to 0.5 on THUMOS and 0.1 to 0.7 on domain adaptation, as well as by averaging the mAP scores across these thresholds. For each tIoU threshold, the Average Precision (AP) for each action class was determined by matching predicted action segments to ground truth segments; a prediction was considered correct if its tIoU with a ground truth segment exceeded the threshold. Precision and recall were evaluated at various confidence score thresholds, with AP calculated as the area under the resulting precision-recall curve.

3. Results

3.1. Benchmarking and Initial Ablation

We evaluated our proposed model configurations against several baselines to assess the effectiveness of architectural modifications using the THUMOS dataset (Figure 4). Baseline systems included the ActionFormer (I3D+RGB) with its original detection head, as well as a variant where the original 1D convolutional (MLP) detection head was replaced with a KAN. For AdaTAD, we compared the original architecture (VideoMAE + BottleneckAdapter + MLP) with our modified versions and also examined variants in which the adapter module was removed to align with baseline settings.

Across all experiments, replacing the convolution (MLP) decoder with a KAN decoder consistently degraded performance. For example, the KAN-equipped ActionTransformer achieved an average mAP of 68.7%, compared to higher scores achieved by MLP-based counterparts. In contrast, our modified AdaTAD models with MLP decoders achieved substantially stronger results,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

with average mAPs ranging from 74% to 75%, outperforming the baselines, including the original AdaDAT. Among adapter designs, both the proposed InvertedConvNeXtAdapter (83.3% mAP) and GRNConvNeXtAdapter (83.5% mAP) slightly outperformed the original BottleneckAdapter (82.5% mAP), indicating that the new adapters provide modest yet consistent gains over the established design. Since ICNA and GRNNA have similar performance, to facilitate further evaluation, we will omit InvertedConvNeXtAdapter in section 3.2.

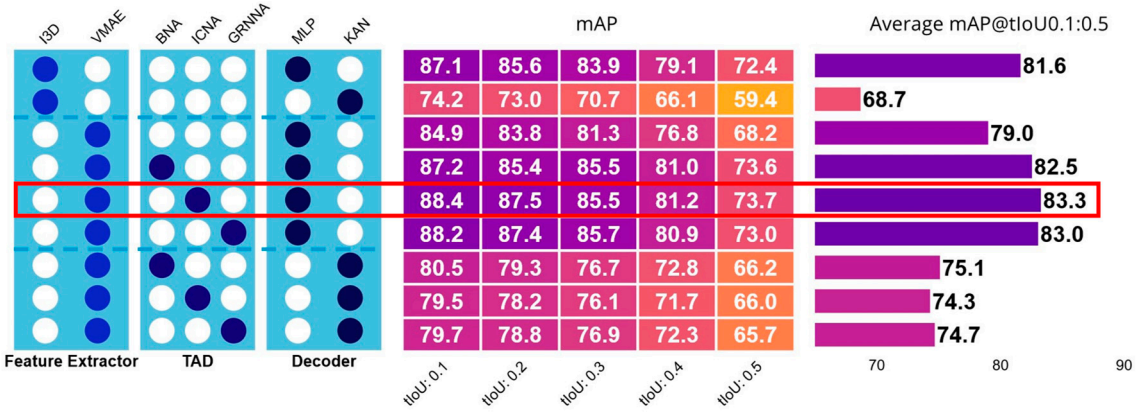

Figure 4. Performance of ablated modified AdaTAD compared with baseline models in the benchmark THUMOS dataset. BNA: BottleneckAdapter; GRNNA: GRNConvNeXtAdapter; I3D: Inflated 3D ConvNet; ICNA: InvertedConvNeXtAdapter; KAN: Kolmogorov-Arnold Network; mAP: mean average precision; MLP: multilayer perceptron via 1D convolution; tIoU: Intersection over Union; VMAE: Video Masked Autoencoder.

3.2. Domain Adptation and Model Performance

Overall, VideoMAE substantially outperformed I3D as the feature extractor, with average mAP improving from 61.55% (I3D + MLP + RGB baseline) to 82.50% when paired with the GRNConvNeXtAdapter and boundary regression. Among adapter designs, the GRNConvNeXtAdapter yielded the highest performance (82.50%), consistently surpassing the BottleneckAdapter (81.65%), Adapter+ (81.91%), and Compacter (77.76%). For the decoder, the Kolmogorov–Arnold Network (KAN) offered moderate gains over the baseline MLP in domain-specific settings, improving the average from 81.27% (MLP) to 82.50% (KAN).

The RGD substitution achieved 82.63% average mAP, exceeding RGB (81.72%), RDB (81.27%), DGB (81.50%), and both early and late fusion baselines (78.60–81.83%). Although the average gain over RGB is modest (~0.9 pp), it was consistent across thresholds, with greater benefit at stricter tIoUs (0.5–0.7). Importantly, our dataset spanned two hostel environments with differing ambient illumination, demonstrating robustness across real-world variability without environment-specific retraining. Compared with early and late fusion, RGD matched or exceeded performance while avoiding dual-stream complexity and additional GPU memory cost.

In terms of input modality, substituting the noisy blue channel with depth (RGD) produced superior mAP (82.63%) compared to RGB (81.72%), RDB (81.27%), DGB (81.50%), or late/early RGBD fusion (78.60–81.83%). For regression strategies, boundary-based regression consistently outperformed centerness, achieving 82.50% compared to 81.27%. Finally, for patch embeddings, sinusoidal positional encoding yielded higher average mAP (82.50%) than rotary embeddings (RoPE, 81.83%).

Taken together, the optimal configuration comprised VideoMAE + GRNConvNeXtAdapter + KAN decoder + RGD input + boundary regression + sinusoidal embedding, achieving an average mAP of 82.50%, representing a clear improvement over the baseline I3D + RGB + MLP model (61.55%).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

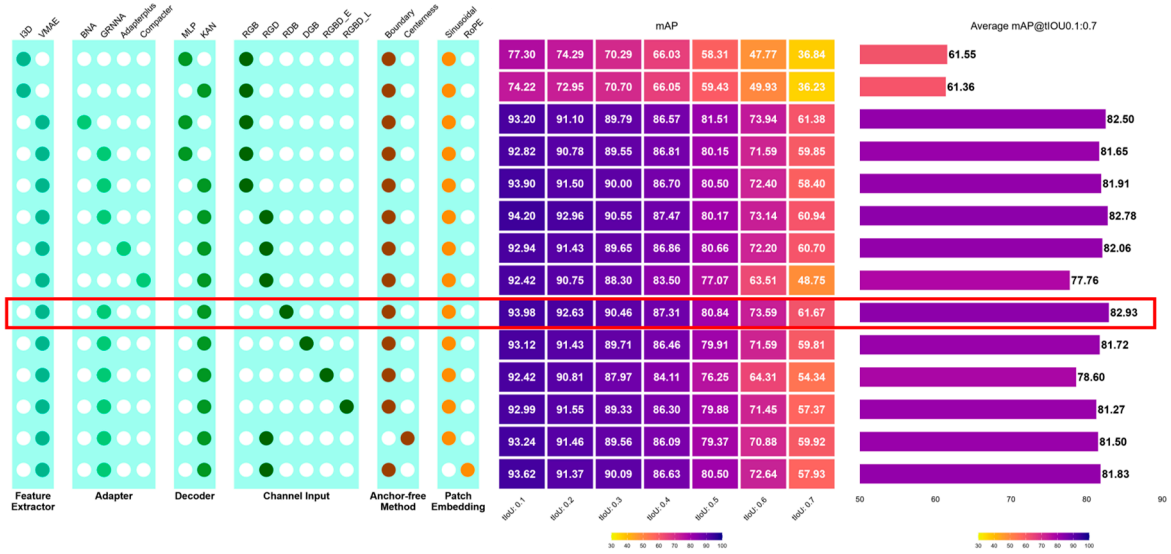

Figure 5. Performance of ablated modified AdaTAD compared with baseline models in the swallowing dataset. BNA: BottleneckAdapter; GRNNA: GRNConvNeXtAdapter; I3D: Inflated 3D ConvNet; ICNA: InvertedConvNeXtAdapter; KAN: Kolmogorov-Arnold Network; mAP: mean average precision; MLP: multilayer perceptron; TFA: temporal feature adapter; tIoU: temporal Intersection over Union; VMAE: Video Masked Autoencoder. RGBD_E: RGB-D with early fusion; RGBD_L: RGB-D with late fusion.

3.3. Adapter Fine-Tuning

After identifying the GRNConvNeXtAdapter as the most effective temporal adapter, we performed a systematic fine-tuning to determine its optimal depth, MLP scaling factor, and the use of residual connections. Figure 6 reports the detailed performance across temporal IoU thresholds (0.1–0.7). Several key trends emerged.

First, increasing the adapter depth improved mAP up to 4 layers, with average performance peaking at 83.25% (Depth = 4, Scale = 2, Residual = No). Beyond 4 layers, deeper settings (6 layers) generally reduced mAP to approximately 81–82%, suggesting over-parameterization and optimization instability. Second, enlarging the scaling factor of the MLP improved representational capacity: mAP rose from 82.53% (Scale = 0.5) to 83.25% (Scale = 2), while further expansion to 4× led to reduced or inconsistent gains (≈82%). Third, enabling residual connections enhanced robustness at stricter thresholds (tIoU ≥ 0.6), yielding up to 74.47% at tIoU 0.6 and 61.86% at tIoU 0.7 (Depth = 4, Scale = 4, Residual = Yes), compared to 72.05% and 59.07% under the corresponding non-residual setting.

We evaluated the Kolmogorov–Arnold Network (KAN) decoder with Chebyshev degree $D = 3$. Compared to the MLP head, KAN offered small but consistent gains on the swallowing dataset (82.50% vs. 81.27% average mAP). The trade-off was modestly higher runtime and memory footprint, though still feasible for research-scale inference. Extended experiments varying D showed stable results, with $D = 3$ providing a good balance between accuracy and efficiency.

Overall, the best-performing configuration was achieved with Depth = 4, Scale = 2, with or without residuals, reaching an average of 83.25% (No residual) and 82.95% (With residual), outperforming both shallower or deeper variants as well as alternative adapters such as BottleneckAdapter, Adapter+, and Compacter.

We systematically evaluated the impact of different feature extractors, adapter modules, decoders, input channel configurations, regression heads, and patch embedding strategies on swallowing event localization. Subsequent to the best model architectural combination, we further fine-tune the adapter configurations based on depth, scaling factor, and addition of residual connection. The detailed results are summarized in Figure 5.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

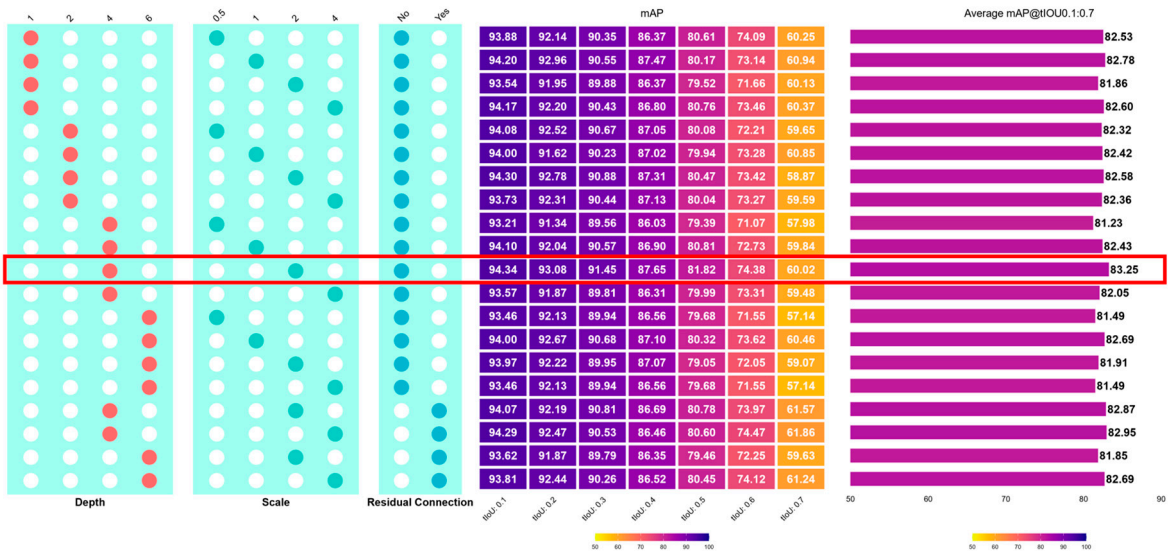

Figure 6. Performance of ablated fine-tuned GRNConvNeXtAdapter **compared with modified model in the swallowing dataset**. mAP: mean average precision; tIoU: temporal Intersection over Union;.

4. Discussion

We present a novel end-to-end RGB-D framework for swallowing event localization that addresses the limitations of purely RGB-based methods. Our core contributions include the proposal of GRNConvNeXtAdapter, which outperforms other existing adapters, and the RGD input strategy, which replaces the noisy blue channel with depth information to capture subtler 3D spatial cues critical for swallowing detection. Experimental results demonstrated significant performance improvements resulting from both the adapter enhancement and the use of RGB-D input. While the KAN decoder did not show any advantage over the existing decoder on the benchmark THUMOS dataset, it outperformed in our proprietary dataset, highlighting its potential in domain-specific applications.

The superior performance of the GRNConvNeXtAdapter can be attributed to its integration of the Global Response Normalization (GRN) layer from ConvNeXt V2, which addresses a critical architectural limitation compared to the BottleneckAdapter. While both adapters enhance temporal feature adaptation, the GRN layer in GRNConvNeXtAdapter employs a three-step mechanism—global L2-norm feature aggregation, divisive normalization for inter-channel competition, and feature calibration—that effectively prevents feature collapse and promotes channel diversity [30]. This mechanism enables adaptive re-calibration of feature channels based on their global response, making them more discriminative for complex temporal patterns inherent in swallowing event detection. Additionally, the GRNConvNeXtAdapter benefits from the larger 7×1 depth-wise convolution kernel that captures longer-range temporal dependencies more effectively than traditional bottleneck designs [30]. The combination of enhanced inter-channel feature competition through GRN and improved temporal modeling creates a more robust architecture that learns balanced and informative representations across all feature channels, leading to superior swallowing event localization performance without additional computational overhead.

The superior performance of RGD input over other channel combinations can be attributed to the replacement of the inherently noisy blue channel with highly informative depth data that captures critical three-dimensional spatial information essential for swallowing event detection. This approach is inspired by the principles of video photoplethysmography (vPPG), where research consistently demonstrates that the red and green channel provides the strongest plethysmography signals due to haemoglobin absorption characteristics, while the blue channel is typically the most susceptible to noise and provides minimal physiological information [36]. Removing the red channel resulted in a degradation of performance, which aligns with the findings of video photoplethysmography (vPPG)

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

theory. Similarly, in digital imaging systems, the blue channel suffers from lower sensor sensitivity and greater light scattering [37,38], making it a prime candidate for replacement. By substituting the blue channel with depth information, the RGD approach provides the model with spatial cues that enable accurate capture of soft tissue deformations and movement patterns in the neck region, critical features that are often invisible or poorly represented in conventional 2D RGB imagery.

Although the improvement of RGD over others is numerically small, it is consistent and particularly evident at stricter thresholds. This non-invasive inspection modality reveals structural information invisible to conventional imaging. Depth complements RGB by exposing geometric cues (soft tissue deformation, laryngeal elevation) not reliably captured by color channels alone.

Initial benchmarks on the THUMOS dataset showed that KANs offered no clear advantage over MLP, likely due to KANs' higher computational overhead and the relatively simple or general nature of video data in THUMOS. However, in our swallowing event localization experiments using specialized RGBD data, KANs outperformed MLPs, likely because the complex feature space of RGBD data better leverages KANs' advanced representational capacity. The compositional structure of KANs and their flexible, learnable activations appear especially adept at capturing the subtle, localized non-linear dynamics of swallowing events, making their theoretical strengths practically beneficial in this domain despite computational challenges.

Directly training a fully end-to-end RGB-D model is hindered by the mismatched characteristics of RGB and depth data—different resolutions, noise profiles, and bit depths—and the enormous computational and memory demands of processing two high-dimensional video streams simultaneously. As a result, many workflows resort to a pseudo end-to-end pipeline, with RGB-D features extracted or pre-trained separately before being fused in a downstream detector, which prevents joint optimization of multimodal representations. In this study, we followed the parameter-efficient fine-tuning strategy of AdaTAD by freezing a pre-trained backbone and inserting lightweight adapters that we modify, thereby preserving learned knowledge, drastically reducing resource requirements, and enabling effective adaptation to RGD swallowing event localization without the prohibitive costs of full end-to-end retraining. Additionally, we attempted to accommodate all RGB-D channels by early and late fusion, despite the performance is worse than the RGD combination.

In our study, we evaluated model performance using mAP at tIoU thresholds ranging from 0.1 to 0.7, align with the conventional range of 0.3 to 0.7 used in some other localization studies [39,40]. This choice was deliberate, catering both temporal accuracy and precise temporal boundary delineation. Our decision to modify the AdaTAD architecture was driven by its demonstrated strength in achieving high temporal accuracy, which is well-suited for detecting the core dynamics of swallowing events, such as hyoid bone elevation, typically occurring in the central portion of the video sequence. In contrast, models like TALLFormer [41], which emphasize precise temporal boundary prediction, often demand significantly greater computational resources, making them less practical for our application. Given that the primary features of swallowing, such as laryngeal elevation and hyoid movement, are most prominent in the middle of the event, precise boundary localization is less critical.

Annotation transparency is central to clinical adoption. We defined swallowing onset as the first frame of rapid laryngeal elevation and offset as the return to baseline hyoid position. An occupational therapist performed primary annotations, which were checked by research staff, and disagreements were resolved by consensus. While formal inter-annotator agreement statistics were not computed, a limitation of the present work, the protocol ensured clear onset/offset identification.

Swallowing is inherently biomechanical, governed by coordinated soft tissue and skeletal dynamics. Future research should integrate finite-element (FEM) simulations to provide motion priors or synthetic sequences that improve interpretability and generalization. Such simulations could align predicted trajectories with physiological expectations, thereby enhancing both performance and clinical credibility.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

The choice of preprocessing also warrants consideration. Depth clipping to 1.0 m and rescaling to 8-bit were pragmatic design decisions that standardized dynamic range and suppressed irrelevant background disparity. Empirically, these steps preserved salient spatial features, with negligible observed impact on performance. Cross-validation was not conducted due to compute limitations and the substantial annotation effort required for our dataset; instead, an 80/20 stratified split was employed, consistent with standard practice in temporal action detection.

Our experiments also provide insight into decoder trade-offs. The Kolmogorov–Arnold Network (KAN) with Chebyshev degree $D = 3$ yielded gains over MLP on the swallowing dataset, albeit at modest runtime and memory overheads. While KANs did not outperform MLPs in general-purpose video benchmarks, their compositional flexibility appears advantageous in specialized biomedical settings. Exploring alternative KAN variants, such as Fourier- or spline-based formulations, may further expand their utility.

Finally, qualitative analysis highlights that while temporal saliency visualization methods (e.g., Grad-CAM) are less informative for dense action localization tasks, case-level inspection remains valuable. We observed scenarios where RGD corrected errors made by RGB alone, particularly under variable lighting and subtle tissue motion. These examples reinforce the role of depth in providing robust cues that generalize across naturalistic monitoring conditions. In home-based deployment scenarios, however, numerous uncontrolled factors, such as ambient lighting variability and sensor placement, can degrade performance. A generalized AI model is therefore needed to enhance robustness across heterogeneous environments. Optimization of medical devices is crucial for the efficiency of healthcare treatments [42], and future research should focus on building adaptive systems capable of handling these dynamic, real-world conditions.

There were several limitations to be noted. Due to strict computational power constraints, batch normalization was applied with an extremely small batch size of two, which is suboptimal. This is due to the This setting results in unstable batch statistics, introducing noise and hindering effective training and generalization [43]. A promising remedy in future work is video-specific patch/token compression (e.g., token merging or learned patch compression) to reduce visual tokens per frame, lowering memory usage and enabling larger BN batches or normalization methods less sensitive to batch size. Additionally, our use of KAN was limited to a Chebyshev polynomial-based variant, yet various KAN architectures exist [44], each with strengths tailored to different data characteristics (e.g., Fourier KANs [45], B-spline KANs [33]). Currently, selecting the optimal KAN type lacks systematic guidelines and often relies on empirical testing, increasing computational demands [44]. Advancing automated or adaptive KAN selection strategies or developing theoretical insights into basis function suitability for specific tasks represents a crucial direction for future research.

5. Conclusions

This study introduced an end-to-end RGB–D framework for swallowing event localization that overcomes key limitations of RGB-only approaches in dysphagia monitoring. We enhanced the AdaTAD backbone through three innovations: (1) an RGD input strategy replacing the noisy blue channel with depth information to capture subtle soft-tissue deformations, (2) a GRNConvNeXtAdapter incorporating Global Response Normalization for improved temporal feature adaptation, (3) a Kolmogorov–Arnold Network (KAN) decoder leveraging Chebyshev polynomials for non-linear temporal pattern recognition, and (4) a boundary regression head and sinusoidal patch embedding method to further optimize model precision

Comprehensive evaluation on our proprietary swallowing dataset (641 clips, 3,153 events) demonstrated substantial improvements. The optimized configuration (VideoMAE + GRNConvNeXtAdapter + KAN + RGD + boundary regression + sinusoidal embedding) achieved an average mAP of 83.25%, compared with 61.55% for the baseline I3D + RGB + MLP model. Ablation studies confirmed that each component—depth substitution, adapter design, regression strategy, and embedding choice—contributed incrementally to the overall performance.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

These findings establish the feasibility of precise, automated temporal localization of swallowing events in continuous video streams without manual segmentation. By combining architectural innovations with a principled modality substitution, this framework advances the development of non-invasive, real-world dysphagia screening tools, supporting future deployment in clinical and home-care monitoring.

Author Contributions: Conceptualization: Duo Wai-Chi Wong and James Chung-Wai Cheung; Data curation, Derek Ka-Hei Lai, Zi-An Zhao, Andy Yiu-Chau Tam, Jing Li and Jason Zhi-Shen Zhang; Formal analysis, Derek Ka-Hei Lai, Zi-An Zhao, Andy Yiu-Chau Tam, Jing Li and Jason Zhi-Shen Zhang; Funding acquisition, James Chung-Wai Cheung; Investigation, Derek Ka-Hei Lai, Zi-An Zhao, Andy Yiu-Chau Tam, Jing Li and Jason Zhi-Shen Zhang; Methodology, Duo Wai-Chi Wong and James Chung-Wai Cheung; Project administration, Duo Wai-Chi Wong and James Chung-Wai Cheung; Software, Derek Ka-Hei Lai, Zi-An Zhao and Andy Yiu-Chau Tam; Supervision, Duo Wai-Chi Wong and James Chung-Wai Cheung; Validation, Derek Ka-Hei Lai and Andy Yiu-Chau Tam; Visualization, Zi-An Zhao and Andy Yiu-Chau Tam; Writing – original draft, Derek Ka-Hei Lai and Zi-An Zhao; Writing – review & editing, Duo Wai-Chi Wong and James Chung-Wai Cheung. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Health and Medical Research Fund from the Health Bureau, Hong Kong (reference number: 19200461 and 21221871).

Institutional Review Board Statement: This study was approved by The Hong Kong Polytechnic University Institutional Review Board (Reference Number: HSEARS20230302009)

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Due to privacy concerns and ethical considerations, the RGB-D data containing identifiable facial information of patients cannot be publicly disclosed.

Acknowledgments: We thank Kowloon Home for the Aged Blind from The Hong Kong Society for the Blind, Kei Oi Neighborhood Elderly Centre and Li Ka Shing Care and Attention Home for the Elderly from the Hong Kong Sheng Kung Hui for facilitating subject recruitment and facilities to conduct experiment.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Putri, A.R.; Chu, Y.-H.; Chen, R.; Chiang, K.-J.; Banda, K.J.; Liu, D.; Lin, H.-C.; Niu, S.-F.; Chou, K.-R. Prevalence of swallowing disorder in different dementia subtypes among older adults: a meta-analysis. *Age and Ageing* **2024**, *53*, afae037.
2. Altman, K.W.; Yu, G.-P.; Schaefer, S.D. Consequence of dysphagia in the hospitalized patient: impact on prognosis and hospital resources. *Archives of Otolaryngology–Head & Neck Surgery* **2010**, *136*, 784-789.
3. Lear, C.S.; Flanagan Jr, J.; Moorrees, C. The frequency of deglutition in man. *Archives of oral biology* **1965**, *10*, 83-IN15.
4. Kramarow, E.; Warner, M.; Chen, L.-H. Food-related choking deaths among the elderly. *Injury prevention* **2014**, *20*, 200-203.
5. De Sire, A.; Ferrillo, M.; Lippi, L.; Agostini, F.; de Sire, R.; Ferrara, P.E.; Raguso, G.; Riso, S.; Roccuzzo, A.; Ronconi, G. Sarcopenic dysphagia, malnutrition, and oral frailty in elderly: a comprehensive review. *Nutrients* **2022**, *14*, 982.
6. Bhattacharyya, N. The prevalence of dysphagia among adults in the United States. *Otolaryngology--Head and Neck Surgery* **2014**, *151*, 765-769.
7. Wu, C.-P.; Chen, Y.-W.; Wang, M.-J.; Pinelis, E. National trends in admission for aspiration pneumonia in the United States, 2002–2012. *Annals of the American Thoracic Society* **2017**, *14*, 874-879.
8. de Castro, M.A.F.; Dedivitis, R.A.; de Matos, L.L.; Baraúna, J.C.; Kowalski, L.P.; de Carvalho Moura, K.; Partezani, D.H. Endoscopic and videofluoroscopic evaluations of swallowing for dysphagia: a systematic review. *Brazilian Journal of Otorhinolaryngology* **2025**, *91*, 101598.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

9. Feng, H.-Y.; Zhang, P.-P.; Wang, X.-W. Presbyphagia: Dysphagia in the elderly. *World journal of clinical cases* **2023**, *11*, 2363.
10. Lai, D.K.-H.; Cheng, E.S.-W.; Lim, H.-J.; So, B.P.-H.; Lam, W.-K.; Cheung, D.S.K.; Wong, D.W.-C.; Cheung, J.C.-W. Computer-aided screening of aspiration risks in dysphagia with wearable technology: a Systematic Review and meta-analysis on test accuracy. *Frontiers in bioengineering and biotechnology* **2023**, *11*, 1205009.
11. Wong, D.W.-C.; Wang, J.; Cheung, S.M.-Y.; Lai, D.K.-H.; Chiu, A.T.-S.; Pu, D.; Cheung, J.C.-W.; Kwok, T.C.-Y. Current Technological Advances in Dysphagia Screening: Systematic Scoping Review. *Journal of Medical Internet Research* **2025**, *27*, e65551.
12. So, B.P.-H.; Chan, T.T.-C.; Liu, L.; Yip, C.C.-K.; Lim, H.-J.; Lam, W.-K.; Wong, D.W.-C.; Cheung, D.S.K.; Cheung, J.C.-W. Swallow detection with acoustics and accelerometric-based wearable technology: a scoping review. *International journal of environmental research and public health* **2022**, *20*, 170.
13. Yao, K.-Y.; Lai, D.K.-H.; Lim, H.-J.; So, B.P.-H.; Chan, A.C.-H.; Yip, P.Y.-M.; Wong, D.W.-C.; Dai, B.; Zhao, X.; Wong, S.H.D.; et al. 2H-MoS2 lubrication-enhanced MWCNT nanocomposite for subtle bio-motion piezoresistive detection with deep learning integration. *Materials & Design* **2025**, *253*, 113861, doi:https://doi.org/10.1016/j.matdes.2025.113861.
14. Sakai, K.; Gilmour, S.; Hoshino, E.; Nakayama, E.; Momosaki, R.; Sakata, N.; Yoneoka, D. A machine learning-based screening test for sarcopenic dysphagia using image recognition. *Nutrients* **2021**, *13*, 4009.
15. Yamamoto, Y.; Sato, H.; Kanada, H.; Iwashita, Y.; Hashiguchi, M.; Yamasaki, Y. Relationship between lip motion detected with a compact 3D camera and swallowing dynamics during bolus flow swallowing in Japanese elderly men. *Journal of Oral Rehabilitation* **2020**, *47*, 449-459.
16. Lai, D.K.-H.; Cheng, E.S.-W.; So, B.P.-H.; Mao, Y.-J.; Cheung, S.M.-Y.; Cheung, D.S.K.; Wong, D.W.-C.; Cheung, J.C.-W. Transformer Models and Convolutional Networks with Different Activation Functions for Swallow Classification Using Depth Video Data. *Mathematics* **2023**, *11*, 3081.
17. Lim, H.-J.; Lai, D.K.-H.; So, B.P.-H.; Yip, C.C.-K.; Cheung, D.S.K.; Cheung, J.C.-W.; Wong, D.W.-C. A comprehensive assessment protocol for swallowing (CAPS): Paving the way towards computer-aided dysphagia screening. *International journal of environmental research and public health* **2023**, *20*, 2998.
18. Hu, K.; Shen, C.; Wang, T.; Xu, K.; Xia, Q.; Xia, M.; Cai, C. Overview of temporal action detection based on deep learning. *Artificial Intelligence Review* **2024**, *57*, 26.
19. Shou, Z.; Chan, J.; Zareian, A.; Miyazawa, K.; Chang, S.-F. CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, USA, 21 - 26 Jul, 2017; pp. 5734-5743.
20. Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; Lin, D. Temporal action detection with structured segment networks. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22- 29 Oct, 2017; pp. 2914-2923.
21. Zhang, C.; Wu, J.; Li, Y. ActionFormer: Localizing Moments of Actions with Transformers. *ArXiv preprint* **2022**, arXiv:2202.07925, doi:10.48550/arXiv.2202.07925.
22. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 21 - 26 Jul, 2017; pp. 6299-6308.
23. Liu, S.; Zhang, C.-L.; Zhao, C.; Ghanem, B. End-to-end temporal action detection with 1b parameters across 1000 frames. In Proceedings of the the IEEE/CVF conference on computer vision and pattern recognition, Seattle, USA, 16 - 22 Jun, 2024; pp. 18591-18601.
24. Tong, Z.; Song, Y.; Wang, J.; Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* **2022**, *35*, 10078-10093.
25. Han, Z.; Gao, C.; Liu, J.; Zhang, J.; Zhang, S.Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608* **2024**.
26. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P. The kinetics human action video dataset. *arXiv preprint* **2017**, arXiv:1705.06950.
27. Lin, Z.; Geng, S.; Zhang, R.; Gao, P.; De Melo, G.; Wang, X.; Dai, J.; Qiao, Y.; Li, H. Frozen clip models are efficient video learners. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 25 - 27 Oct, 2022; pp. 388-404.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2025 doi:10.20944/preprints202510.0294.v1

28. Pratticò, D.; Laganà, F.; Oliva, G.; Fiorillo, A.S.; Pullano, S.A.; Calcagno, S.; Carlo, D.D.; Foresta, F.L. Integration of LSTM and U-Net Models for Monitoring Electrical Absorption With a System of Sensors and Electronic Circuits. *IEEE Transactions on Instrumentation and Measurement* **2025**, *74*, 1-11, doi:10.1109/TIM.2025.3573363.
29. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, USA, 18 - 24 Jun, 2022; pp. 11976-11986.
30. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the the IEEE/CVF conference on computer vision and pattern recognition, Vancouver, Canada, 17 - 24 Jun, 2023; pp. 16133-16142.
31. Steitz, J.-M.O.; Roth, S. Adapters Strike Back. **2024**, arXiv:2406.06820, doi:10.48550/arXiv.2406.06820.
32. Karimi Mahabadi, R.; Henderson, J.; Ruder, S. Compacter: Efficient Low-Rank Hypercomplex Adapter Layers. **2021**, arXiv:2106.04647, doi:10.48550/arXiv.2106.04647.
33. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. Kan: Kolmogorov-arnold networks. *arXiv preprint* **2024**, arXiv:2404.19756.
34. SS, S.; AR, K.; KP, A. Chebyshev polynomial-based kolmogorov-arnold networks: An efficient architecture for nonlinear function approximation. *arXiv preprint* **2024**, arXiv:2405.07200.
35. Idrees, H.; Zamir, A.R.; Jiang, Y.-G.; Gorban, A.; Laptev, I.; Sukthankar, R.; Shah, M. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding* **2017**, *155*, 1-23.
36. Kumar, M.; Veeraraghavan, A.; Sabharwal, A. DistancePPG: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express* **2015**, *6*, 1565-1588.
37. Wyatt, P.J. Differential light scattering and the measurement of molecules and nanoparticles: A review. *Analytica Chimica Acta: X* **2021**, *7*, 100070.
38. Liu, P.T.; Ruan, D.B.; Yeh, X.Y.; Chiu, Y.C.; Zheng, G.T.; Sze, S.M. Highly responsive blue light sensor with amorphous indium-zinc-oxide thin-film transistor based architecture. *Scientific reports* **2018**, *8*, 8153.
39. Liberatori, B.; Conti, A.; Rota, P.; Wang, Y.; Ricci, E. Test-time zero-shot temporal action localization. In Proceedings of the the IEEE/CVF conference on computer vision and pattern recognition, Seattle, USA, 17 - 21 Jun, 2024; pp. 18720-18729.
40. Reka, A.; Borza, D.L.; Reilly, D.; Balazia, M.; Bremond, F. Introducing Gating and Context into Temporal Action Detection. In Proceedings of the European Conference on Computer Vision, MiCo Milano, Italy, 8 - 13 Sep, 2024; pp. 322-334.
41. Cheng, F.; Bertasius, G. TALLFormer: Temporal Action Localization with a Long-memory Transformer. *ArXiv preprint* **2022**, arXiv:2204.01680, doi:10.48550/arXiv.2204.01680.
42. Laganà, F.; Faccì, A.R. Parametric optimisation of a pulmonary ventilator using the Taguchi method. *Journal of Electrical Engineering* **2025**, *76*, 265-274, doi:10.2478/jee-2025-0027.
43. Keskar, N.S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; Tang, P.T.P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint* **2016**, arXiv:1609.04836.
44. Somvanshi, S.; Javed, S.A.; Islam, M.M.; Pandit, D.; Das, S. A survey on kolmogorov-arnold network. *ACM Computing Surveys* **2025**, doi: 10.1145/3743128.
45. Zhang, J.; Fan, Y.; Cai, K.; Wang, K. Kolmogorov-Arnold Fourier Networks. *arXiv preprint* **2025**, arXiv:2502.06018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.