2by2: Weakly-Supervised Learning for Global Action Segmentation

Elena Bueno-Benito[0009−0006−7566−9771] and Mariella Dimiccoli[0000−0002−2669−400X]

Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens i Artigas 4-6, 08028 Barcelona, Spain
{ebueno, mdimiccoli}@iri.upc.edu

Abstract. This paper presents a simple yet effective approach for the poorly investigated task of global action segmentation, aiming at grouping frames capturing the same action across videos of different activities. Unlike the case of videos depicting all the same activity, the temporal order of actions is not roughly shared among all videos, making the task even more challenging. We propose to use activity labels to learn, in a weakly-supervised fashion, action representations suitable for global action segmentation. For this purpose, we introduce a triadic learning approach for video pairs, to ensure intra-video action discrimination, as well as inter-video and inter-activity action association. For the backbone architecture, we use a Siamese network based on sparse transformers that takes as input video pairs and determine whether they belong to the same activity. The proposed approach is validated on two challenging benchmark datasets: Breakfast and YouTube Instructions, outperforming state-of-the-art methods.

Keywords: Temporal Action Segmentation · Weakly-Supervised Learning · Video Alignment.

1 Introduction

Action segmentation, the task of classifying each frame of an untrimmed video plays a fundamental role in various applications such as video surveillance, sports analysis, and content-based video retrieval [21, 50]. Recently, this task has received significant attention from the research community. The most reliable approaches for action segmentation are fully supervised methods, which require expensive data annotations [5, 6, 19, 27, 32, 48]. The need for more scalable and practical solutions has led to an increasing interest in developing weakly-supervised [9, 30, 31, 33, 40, 46, 49] and unsupervised techniques [7, 11, 12, 14, 23, 24, 26, 28, 35, 37, 42, 43, 45, 47].

Weakly-supervised methods learn to partition videos into action segments using only transcript annotations for each video, typically in the form of actions transcripts (ordered lists of actions) or action sets (unique actions derived from narrations, captions or meta-tags) [31, 40, 46, 49]. This weakly-supervised paradigm contrasts with unsupervised methods, broadly categorized into three

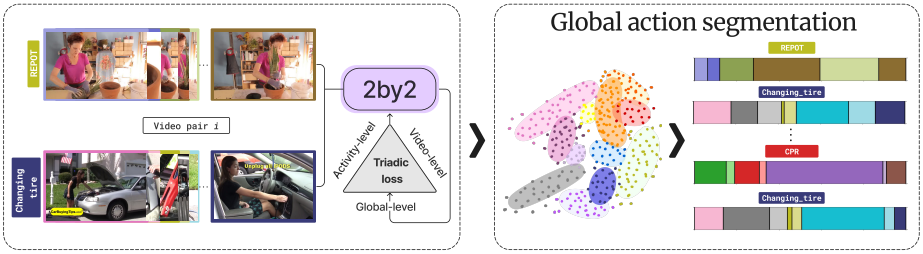

Fig. 1: Our approach compares video pairs through a Siamese network by using binary labels indicating if the videos belong to the same activity or not. We propose a triadic loss function modelling intra-video discrimination, inter-video and inter-activity associations for clustering actions across videos of different activities.

types, depending on the matching objective [13]: video-level, activity-level, and global-level. Video-level segmentation methods aim to segment a single video sequence into distinct actions without considering the relationships between actions in different videos [7, 16, 28, 35, 47]. While they can be effective for practical applications requiring to segment isolated videos one by one, they fail to generalize actions across different videos. Instead, activity-level segmentation methods focus on matching actions across videos that depict the same complex activity [14, 23, 24, 26, 42, 46]. These methods generally perform poorly at video-level unless temporal smoothing within segments is explicitly modelled. In addition, as they assume or estimate a transcript for each video or set of videos belonging to the same activity, their generalization ability to other activities is hampered. Only Ding *et al.*[14] directly addressed global-level segmentation railing on complex activity labels to help discover the constituent actions; however, they do not explicitly model the alignment of actions across videos of the same activity.

In this paper, we propose a strategy to discover actions across various complex activity videos, offering a broader and more generalized understanding of actions. Our approach does not require knowledge of video transcripts, but only binary labels indicating whether each pair of videos belongs to the same activity. Therefore, as a weakly-supervised method, it occupies a unique position in the spectrum of action segmentation methods.

Our solution, depicted in Figure 1, aims to enhance the clustering of actions in videos on a global scale through the implementation of a Siamese network based on transformers. This network is designed to address the task of determining whether two videos depict the same activity. Instead of using a standard cross-entropy loss, we propose a triadic loss function capturing the temporal dynamics within individual videos, between similar videos, and across various activities. Our contributions are as follows:

1. We propose a novel weakly-supervised framework for the task of global action segmentation that relies on binary activity labels to discover action clusters across videos of different activities.

2. We introduce a transformer-based Siamese architecture, that takes input pairs of videos, determines if they belong to the same activity or not and aligns them temporally if predicts that they depict the same activity.
3. We introduce a triadic loss function that models intra-video action discrimination at the video-level, inter-video and inter-activity action associations at activity and global-level respectively, for robust action understanding.
4. We achieve state-of-the-art results on the *Breakfast (BF)* and *Inria Instructional Videos (YTI)* benchmark datasets, demonstrating the method's effectiveness and generalization ability across activities.

2 Related work

2.1 Action Segmentation

For a comprehensive and recent survey on temporal action segmentation tasks, readers are referred to [13].

Supervised Action Segmentation. Supervised approaches have seen significant advancements over recent years [5, 6, 19, 27, 32, 48]. Recently, UVAST [6] integrates fully and timestamp-supervised learning paradigms via sequence-to-sequence translation. This method refines predictions by aligning frame labels with predicted action sequences. LTContext [5] iterates between windowed local attention and sparse long-term context attention, effectively balancing computational complexity and segmentation accuracy. Lastly, FACT [32] performs temporal modelling at both frame-level and action-level, facilitating bidirectional information transfer and iterative feature refinement. However, being fully supervised, all these methods are not scalable and not suited for real applications.

Weakly-Supervised Action Segmentation. Weakly-supervised techniques have been developed to reduce the need for large annotated datasets. These approaches typically learn to partition a video into several action segments from training videos only using transcripts or other human-generated information to generate pseudo-labels for training [30, 31, 33, 40, 46, 49]. Transcripts have been shown to outperform action set-based methods, while timestamp-based approaches achieve the best results. This suggests that higher levels of supervision generally lead to better performance. In recent years, DP-DTW [9] has advanced weakly-supervised segmentation by training class-specific discriminative action prototypes. This method represents videos by concatenating prototypes based on transcripts and improves inter-class distinction through discriminative losses. Some methods leverage machine learning models to infer video segments, such as TASL [30]. Recently, more efficient alignment-free methods have been proposed. MuCon [40] learns from the mutual consistency between two forms of segmentation: framewise classification and category/length pairs. POC [31] introduces a loss function to ensure the output order of any two actions aligns with the transcript. Conversely, ATBA [46] propose an approach that incorporates alignment by directly localizing action transitions for efficient pseudo-segmentation generation during training, eliminating the need for time-consuming frame-by-frame

alignment. None of these methods explicitly addresses the problem of global action segmentation.

Unsupervised Action Segmentation. Unsupervised approaches have been explored by several studies to eliminate the need for annotations [1, 7, 11, 14, 16, 23, 24, 26, 28, 35–37, 42, 43, 45, 47]. As the estimated clusters, lack of semantic labels, the evaluation process requires finding the Hungarian correspondence between the clusters and the actual action classes. The Hungarian matching can be performed for video-level segmentation [1, 7, 16, 28, 35, 36], activity-level segmentation [14, 23, 24, 26, 37, 42, 43, 45, 47], or for a global scope across an entire set of videos [14, 23, 26]. Depending on the hierarchical level used, methods aim to improve segmentation through these correspondences. Unsupervised techniques in action segmentation typically involve a two-step process: first, learning action representations in a self-supervised manner, followed by employing clustering algorithms to perform action segmentation, assuming a prior knowledge of the number of clusters.

In the realm of video-level action segmentation, LSTM+AL [1] introduced a novel self-supervised methodology for real-time action boundary detection. Furthermore, it is worth noticing that clustering approaches based on specific similarity metrics have been relatively under-explored in the field of action segmentation. One such method is TW-FINCH [35], which captures spatio-temporal similarities among video frames. This employs a temporally weighted hierarchical clustering algorithm, grouping video frames without the need for extensive pre-training, as it directly operates on pre-computed features that augment the conventional FINCH approach with temporal considerations [36]. In a similar vein, ABD[16] identifies action boundaries by detecting abrupt change points along the similarity chain between consecutive features.

Action representation learning at the individual video level has also gained interest. TSA [7] proposed a method that focuses on this aspect, employing a shallow neural network trained with a triplet loss and a novel triplet selection strategy to learn action representations. These learned representations can be processed using generic clustering algorithms to obtain segmentation outputs. Lastly, the OTAS framework has emerged, offering an unsupervised boundary detection method that combines global visual features, local interacting features, and human-object relational features, contributing to the evolving landscape of action segmentation techniques [28].

Some approaches at the activity-level leverage the order of scripted activities, emphasizing the minimization of prediction errors, like CTE [23]. Other works combined temporal embedding with visual encoder-decoder pipelines with visual reconstruction loss [43] or with discriminative embedding loss [41]. ASAL [26] explored deep learning architectures, such as ensembles of autoencoders and classification networks that exploit the relationship between actions and activities. CAD [14] introduced a framework that discovers global action prototypes based on high-level activity labels. One notable aspect of these methods is the recognition that actions in task-oriented videos tend to occur in similar temporal contexts. As a result, strong temporal regularization techniques have been developed

to partially obscure visual similarities [23, 37]. Recently, optimal transport has gained popularity in unsupervised learning to generate effective pseudo-labels and train for frame-level action classification. TOT [24] proposed a joint self-supervised representation learning and online clustering approach that directly optimizes unsupervised activity segmentation using video frame clustering as a pretext task. UFSA [42] extends TOT by combining frame and segment-level cues to improve permutation-aware activity segmentation. Furthermore, TOT and UFSA use a Hidden Markov Model (HMM) approach to decode segmentations given a fixed or estimated action order, respectively. In contrast, ASOT [47] proposed a method via optimal transport that yields temporally consistent segmentations without prior knowledge of the action ordering, required by previous approaches. Suitable for both pseudo-labeling and decoding.

Although global-level understanding provides the most comprehensive insight into the relationships between activities and actions in videos, only a few methods have explored training at this level. CAD [14] is the first work to operate at the highest level of global matching. In CTE [23], the methods extended their configuration considering all complex activities. Firstly, the protocol executes a bag-of-words clustering on the videos to divide them into multiple pseudo-activities. Subsequently, they perform action clustering within each pseudo-activity individually. In other words, they apply their action segmentation at the activity level within classes of pseudo-activity. Their approach still does not accommodate potential actions shared between activities. ASAL [26] and CAD [14] present their results aligned with this protocol.

2.2 Video Alignment

Video alignment is a process aimed at synchronizing and matching video sequences for various applications, such as action recognition model creation, behavioural analysis, and multimedia content generation. This field encompasses a range of techniques. Traditionally, methods like Dynamic Time Warping (DTW), Canonical Correlation Analysis, ranking or match-classification objectives, and the differentiable version of DTW, Soft-DTW, have been used to tackle the challenging task of aligning video frames [3, 4, 10, 38] in videos depicting a same action. Recently, LAV [20] have utilized Soft-DTW combined with temporal intra-video contrastive loss to align video frames effectively. Drop-DTW [17], an extension of DTW, introduces a "trash bucket" to the cost matrix, allowing for the classification of background frames and robust alignment in the presence of outliers. VAVA [29] employs optimal transport with a bi-modal Gaussian prior and a virtual frame for unmatched frames.

TCC [18] was the first to introduce cycle-consistency for aligning video frames by maximizing cycle-consistent embeddings between sequences. GTCC [15] extends the TCC approach to manage more complex alignment scenarios. However, most of these techniques were developed for general video alignment or related tasks, and their direct application to unsupervised action segmentation has been never explored so far. In this paper, we propose for the first time to leverage video alignment for action segmentation.

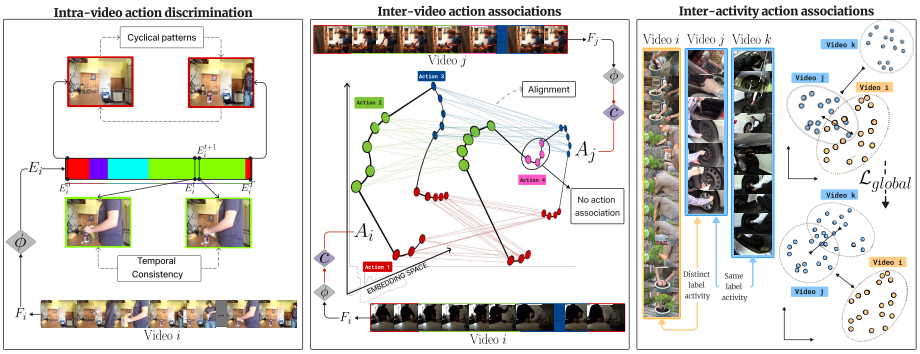

Fig. 2: Overview of the proposed 2by2 framework. The figure illustrates our triadic learning approach: intra-video action discrimination, which enhances cross-temporal consistency within a single video (first box); inter-video action associations, which align action frames among similar videos (second box); and inter-activity action associations, which establish global correspondence between different videos (third box). The red arrows indicate steps specific to the training phase.

3 2by2: Learning Unknown Actions in a Global Manner

This section presents a weakly-supervised, triadic action learning approach for global action segmentation (see Figure 2), aiming at modeling:

(i) *Intra-video action discrimination* (video level): Video frames sharing the same action with their nearest neighbours exhibit temporal consistency. Moreover, actions typically do not occur at the beginning or end of videos. Thus, a video can be interpreted as a cyclic temporal sequence.
(ii) *Inter-video action associations* (activity level): For videos categorized under the same activity, segments within these videos exhibit similarity, facilitating the alignment of actions across them.
(iii) *Inter-activity action associations* (global level): Videos representing different activities that share common actions should be closer in the representational space compared to those that do not share actions.

3.1 Problem Formulation

Given a large set V of complex activity videos from a dataset belonging to C complex activities, each video v_i in V is annotated with a complex activity label $a \in [1, C]$. Our objective is to associate each video frame x_t, with an action label n from N possible actions. These N actions are constituent steps shared among the C complex activities. For each video v_i, we define the feature matrix F_i, where each row F_i^t corresponds to an d-dimensional feature vector at time t in a video v_i. Given the initial features of a video F_i, our objective is to learn a parametric function ϕ that categorizes video frames into the N possible actions, resulting in embeddings E_i, obtained as $E_i = \phi(F_i), \forall v_i \in V$.

3.2 Architecture

To learn ϕ, we propose a Siamese architecture that takes as input pairs (F_i, F_j) for all $v_i, v_j \in V$ with $i \neq j$. This architecture consists of two identical LTContext networks [5], specifically designed to capture long-term temporal dependencies, that work in tandem and compare the similarity between their outputs, denoted as $(E_i,\ E_j)$ at the end.

During training, to ensure that videos sharing the same activity have well-aligned representations, we introduce a context-drop function c, inspired by [15]. This function is designed to handle background and redundant frames by enforcing multi-cycle consistency for alignable embeddings and poor alignment for droppable embeddings. The context-adjusted embeddings are calculated as $A_i = c(E_i), \forall v_i \in V$.

3.3 Triadic Loss

Intra-video Discrimination Loss. The output of ϕ at different stages, denoted as ϕ_s, is used to calculate the loss at video level, enhancing the model's ability to learn fine-grained temporal structures. We incorporate a mean squared error smoothing loss, as introduced by [19] and used in [5, 27, 48]. Considering that actions occurring in an activity video should be temporally contiguous, this loss is applied to the per-frame actions to alleviate over-segmentation. Moreover, we also propose a cyclic variant, based on the assumption (i) described at the beginning of Section 3. Specifically, this variant compares the embeddings at the end of the output sequence with those at the start, across different stages of the feature extraction network ϕ. This is driven by the fact that actions often exhibit cyclical patterns in videos. Mathematically, our video-level loss is defined as follows:

$$\mathcal{L}_{video}(i) = \frac{1}{|S||T+1|}\Big(\sum_{s}\sum_{t}\big|\log\phi_s(F_i^{t+1}) - \log\phi_s(F_i^{t+1})\big| + \big|\log\phi_s(F_i^{T}) - \log\phi_s(F_i^{0})\big|\Big), \quad (1)$$

where T is the total number of frames and S is the number of stages in ϕ in a video i, $\forall v_i \in V$.

Inter-video Associations Loss. For segment-level learning, we adopt the GTCC loss function proposed by [15], denoted as $\mathcal{L}_{activity}$, to synchronize frames of videos depicting the same activity. We utilize context-adjusted embeddings Ai generated by our context-drop function layer c. Specifically, for each pair $v_i, v_j \in V$ of videos, GTCC computes the probability of dropping v_i^t given v_j for all $t \in T$ using the function c. The loss function is defined as:

$$\mathcal{L}_{GTCC}(v_i|v_j) = \sum_{t}\Big((1 - \mathrm{P}_{\mathrm{drop}}(v_i^t|A_j)) \cdot \mathcal{L}_{multi-cbr} + \frac{\mathrm{P}_{\mathrm{drop}}(v_i^t|A_j)}{\mathcal{L}_{multi-cbr}}\Big), \quad (2)$$

where $\mathcal{L}_{multi-cbr}$ is a multi-cycle back regression loss, and $\mathrm{P}_{\mathrm{drop}}(v_i^t|A_j)$ is the probability of dropping each video frame v_i^t given A_j (refer to [15] for more details). Our activity loss, $\mathcal{L}_{activity}$ is defined as the sum of $GTCC$ loss of v_i given A_j and vice-versa. This loss leverages the principle of Temporal Cycle Consistency (TCC) [18], ensuring that corresponding frames in videos with identical action sequences are closely aligned in the feature space. This approach addresses variations in action order, redundant actions, and background frames, thereby enhancing the quality of video representations. To the best of our knowledge, this marks the first application of video alignment for temporal action segmentation.

Inter-activity Associations Loss. We learn the global representation of a video clip by using a contrastive loss. We employ contrastive learning to minimize the distance between videos of the same activity while maximizing the distance between videos of different activities. This ensures that videos depicting the same activity are closer in the feature space than videos that are not. The global contrastive loss has the following formulation:

$$\mathcal{L}_{global}(i,j) = (1-y)\cdot d(E_i,E_j) + y\cdot \max(0, m - d(E_i,E_j)) \tag{3}$$

where $d(E_i,E_j)$ denotes the distance between the representations E_i and E_j obtained by ϕ, and $y \in \{0,1\}$ is a binary value such that $y=0$, if the two videos belong to the same activity ($a_i = a_j$), and $y=1$, if they belong to different activities ($a_i \neq a_j$). The margin m ensures sufficient separation between videos of different activities. The term $(1-y)\cdot d(E_i,E_j)$ minimizes the distance for videos of the same activity, while $y\cdot\max(0, m-d(E_i,E_j))$ maximizes the distance for videos of different activities by pushing them apart by at least the margin m.

The combined loss function that governs the training for all pair videos $\{v_i, v_j\} \in V$ of our model is formulated as:

$$\mathcal{L}_{\mathrm{train}}(\phi,c)) = \begin{cases} \alpha\mathcal{L}_{\mathrm{global}}(i,j) + (1-\alpha)\mathcal{L}_{\mathrm{activity}}(i,j) \\ \qquad\qquad +\beta(\mathcal{L}_{\mathrm{video}}(i) + \mathcal{L}_{\mathrm{video}}(j)), & \text{if } v_i = v_j \\ \mathcal{L}_{\mathrm{global}}(i,j) + \beta(\mathcal{L}_{\mathrm{video}}(i) + \mathcal{L}_{\mathrm{video}}(j)), & \text{if } v_i \neq v_j \end{cases} \tag{4}$$

where α, and β are hyperparameters that balance the contributions of the global, activity, and video loss components. Incorporating this loss in our model allows us to leverage the weak supervision effectively, making the clustering of video frames more discriminative and improving the overall performance of action segmentation and classification tasks in a global manner.

4 Experimental Setup

Datasets. We present results on two well-known datasets used for temporal action segmentation: **Breakfast Action Dataset (BF)** [22] is one of the largest fully annotated collections available for temporal action segmentation. It includes

1712 videos, featuring 10 activities related to breakfast preparation. These activities are performed by 52 individuals across 18 different kitchens. Each video has an average of 2099 frames. Remarkably, only 7% of the frames are background frames. **Youtube INRIA Instructional Dataset (YTI)** [2] includes 150 instructional videos from YouTube, covering 5 different activities such as changing a car tire, preparing coffee, and performing cardiopulmonary resuscitation (CPR). The videos have an average duration of 2 minutes. A significant challenge with this dataset is the high proportion of background frames, which make up 63.5% of the total frames.

Features. To ensure a fair comparison with related work, we utilized the same input features as recent methods. For the BF dataset, we used the IDT features [44] provided by the authors of CTE [22] and SCT [34]. These features capture motion information by tracking dense points in the video and computing descriptors such as Histogram of Oriented Gradients, Histogram of Optical Flow (HOF)[25], and Motion Boundary Histogram. Additionally, for further comparison in the BF dataset, we employ I3D features [8] extracted from the Inflated 3D ConvNet, which leverages both spatial and temporal convolutions to learn video representation. For the YTI dataset, we use the same features as [2, 14]. These 3000-dimensional feature vectors are formed by concatenating HOF descriptors with features extracted from the VGG16-conv5 layer [39].

Metrics. To evaluate the performance of our temporal action segmentation methods, we employ 1) Mean over Frames (MoF), which calculates the accuracy as the mean percentage of correctly classified frames across all videos, providing a direct indication of overall segmentation performance; 2) F1-Score, which is the harmonic mean of precision and recall, accounting for both false positives and false negatives. Precision is the ratio of correctly predicted action frames to the total predicted action frames, while recall is the ratio of correctly predicted action frames to the total actual action frames; 3) MoF with Background (MoF-BG), which calculates the accuracy considering both action and background frames, essential for understanding how well the segmentation method distinguishes between action and non-action frames, especially given the high proportion of background frames in the YTI dataset. To enable direct comparison, we follow the procedure used in previous work [7, 28, 23, 16, 37], reporting results by removing the ratio ($\tau = 75\%$) of the background frames from the video sequence.

Evaluation Setting. In our study, we adopt the global evaluation methodology proposed by [23]. This methodology involves grouping videos into coherent subsets K and representing them using a bag-of-words (BoW) approach. These representations are then clustered into K' groups of pseudo-activities and K subgroups of actions are inferred. Each video is temporally segmented by assigning each frame to one of the ordered groups using the Viterbi decoder. A background model is introduced to deal with irrelevant segments. Throughout

the results of this work, the inclusion of BoW and Decoding refers to the integration of the aforementioned global inference process, which we will refer to as the *post-processing protocol*.

For evaluation, we perform a Hungarian matching between the inferred clusters and the ground-truth labels to compute the metrics. Specifically, we assume in the case of the Breakfast dataset $K' = 10$ activity clusters with $K = 5$ sub-actions per cluster. Subsequently, we match 50 different sub-action clusters with 48 ground-truth sub-action classes, with frames of the leftover clusters set as background. Finally, we assess the accuracy of the unsupervised learning configuration on the YouTube Instructions dataset, employing $K' = 5$ and $K = 9$, subsequently matching 45 distinct sub-action clusters with 47 ground-truth sub-action classes.

Training Details. To ensure that each video in our training set has at least one pair from the same activity and one pair from a different activity, we construct the training set by including all possible combinations of videos belonging to the same activity. Since segment-level learning requires a strong initialization to align actions between videos, we adopt a two-stage training approach. Initially, the model is trained with global-level and video-level modules using Eq. 1 and 3, respectively. Subsequently, the model is used to initialize the second stage, where it is trained using the full loss function in Eq. 4. In a stratified fashion, we select a subset of pairs from different activities, ensuring an equal number of same-activity and different-activity pairs. Given a large number of combinations, in each epoch, we take a batch including 50% of the dataset of possible pairs for each epoch. Note that each epoch uses a batch size of 32 pairs for the BF dataset and 8 pairs for the YTI dataset. We simultaneously train a 4-layer feed-forward neural network for the drop-context function, c, along with ϕ. To enhance computational efficiency, we down-sample all videos to 256 frames per video by randomly removing frames distributed throughout each video, similar to [24, 42]. This technique reduces frame redundancy and ensures that the frames represent the entire video. We use the same parameters as specified in [5] and [15] for each network. The training process employs the ADAM optimizer, with a learning rate of $2e^{-4}$ and a weight decay of 10^{-4}. For the parameters α and β, we select the values 0.15 and 0.5, respectively.

4.1 Comparative methods.

The method more similar to ours in terms of scope, i.e. global action segmentation, and information used, i.e. activity labels, is CAD [14]. For the sack of completeness, we compute results with a global matching scope of state-of-the-art methods conceived for action segmentation at activity level. These include on the one side unsupervised methods such as ASOT [47], CTE [23] and ASAL [26] that train a network for each activity hence using our same pseudo-labels; on the other side, they include weakly-supervised methods such as ATBA [46] that instead use a transcript for each video, resulting in a much stronger level of supervision.

BF

Supervision	Approach	F	BoW	D	F1	MoF
Unsupervised	CTE [23]	IDT	✓	✓	-	18.5
	ASAL [26]	IDT	✓	✓	-	20.2
	ASOT* [47]	IDT	✓	✓	20.2	21.6
Weak	CAD [14]	IDT	✗	✗	-	10.9
			✓	✗	-	17.7
			✓	✓	-	23.4
	2by2	IDT	✓	✓	**20.6**	**24.6**
Unsupervised	ASOT* [47]	I3D	✓	✓	16.9	18.1
Weak-transcripts	ATBA* [46]	I3D	✓	✓	20.0	17.7
Weak-activity labels	CAD [14]	I3D	✗	✗	-	19.2
	2by2	I3D	✓	✓	**17.5**	**20.7**

YTI

Supervision	Approach	BoW	D	F1	MoF	MoF-BG
Unsupervised	CTE [23]	✓	✓	-	19.4	10.1
	ASOT* [47]	✓	✓	15.26	18.6	9.9
Weak	CAD [14]	✗	✗	12.10	15.7	-
	2by2	✓	✓	**16.53**	**23.6**	**11.4**

Table 1: Action Segmentation results on the BF and YTI datasets by applying the Hungarian matching at global-level. The dash indicates "not reported." (*) denotes results computed by ourselves. "F" denotes the type of features used. "D" indicates the use of Viterbi decoding. Both marks denote evaluation as per [23]. The best results are marked in bold.

5 Results

5.1 Comparison with the State-of-the-art

Breakfast dataset (BF). The results obtained by using the IDT features as input demonstrate a consistent performance improvement over prior methods (refer to left-hand table 1). We achieved a +1.2% improvement in MoF with respect to CAD, highlighting the efficacy of our global training approach with binary labels.

We computed the results at the global level of ASOT [47], by following the evaluation protocol described above. 2by2 outperforms it in terms of MoF by +3% and in terms of F1-score by +0.4%. Similar trends are observed when using I3D features as input. Compared to state-of-the-art methods, the 2by2 framework proves effective regarding MoF and F1-score. ATBA [46] exhibits a higher F1-score but a lower MoF than 2by2, likely due to its use of transcripts for each video, providing stronger supervision with respect to our method but poorer generalization across activities. This could be attributed to the fact that these methods were not specifically designed for global training, highlighting the critical importance of inter-activity learning which is currently lacking in other unsupervised methods.

Inria Instructional Videos (YTI). The performance of our 2by2 framework also shows marked improvements over previous methods on the YTI (refer to right-hand table 1). We achieve an increase in MoF of +4.2% without background and +1.3% with background. This improvement in the F1 score is likely attributed to the non-repetitive nature of actions within activities in this dataset. Our 2by2 framework effectively enhances segmentation accuracy compared to ASOT, the leading unsupervised activity-level segmentation method. Similar to BF, our results underscore the effectiveness of inter-activity training.

YTI

$\mathcal{L}_{\text{video}}$	$\mathcal{L}_{\text{activity}}$	$\mathcal{L}_{\text{global}}$	**MoF**
✓	✓	✓	**23.6**
✗	✓	✓	21.9
✓	✗	✓	22.5
✗	✗	✓	21.8

Base	**23.6**
No k_means init	21.1
No cycled MSE	21.0
No k_means init and cycled	20.4

Table 2: Ablation studies on the YTI dataset, highlighting the importance of the three loss terms, as well as of the concept of temporal cycles and the initialization with k-means.

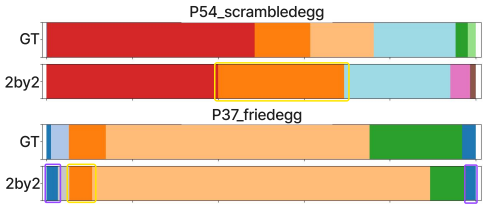

Fig. 3: Examples from BF ("scrambled egg" and "fried egg" activities). Comparison of ground truth (GT) segmentation and our 2by2 framework. 2by2 discovers common action steps across activities (see yellow segments) and captures the cyclic nature of the videos (see purple segments).

Furthermore, leveraging global-level training with CAD, we observe significant improvements of +7.9% in MoF and +4.4% in F1 score.

The observed performance improvements in both datasets are likely due to the framework's ability to identify better shared actions among pseudo-activity classes caused by inaccurate pseudo-labels and the enhanced initialization of the Bag of Words (BoW) model through video alignment.

Qualitative Result. In Fig. 3, we observe examples closely aligning with the ground truth segments, accurately capturing both large and small segments. The enhanced segmentation arises from multi-level processing within our framework. The activity-level component (GTCC) facilitates precise segment alignment, while the global aspect improves activity differentiation and reduces misclassification. At the video level, our framework maintains temporal consistency and cyclic patterns, reducing over-segmentation and enhancing alignment.

5.2 Ablation study

In Table 2, we show the importance of modelling all three levels of learning, by using $\mathcal{L}_{\text{video}}$, $\mathcal{L}_{\text{activity}}$ and $\mathcal{L}_{\text{global}}$. Specifically, we observe that the elimination of the intra-video component significantly impacts our method's performance, highlighting the detrimental effect of relying solely on the global loss. Additionally, since the inter-video component is introduced in the second stage, it becomes clear that robust initialization in the first stage is essential for $\mathcal{L}_{\text{activity}}$ to effectively guide the alignment and segmentation processes. This underscores that the global loss alone in the first stage is insufficient for achieving optimal performance.

Furthermore, we ablate the effect of initializing the activity cluster for the last layer used for $\mathcal{L}_{\text{global}}$ by using k-means instead of random initialization. Additionally, the negative impact of removing the cyclic component from $\mathcal{L}_{\text{video}}$ is evident.

6 Conclusion

This paper introduced 2by2, a novel framework for weakly supervised temporal action segmentation in untrimmed videos encompassing different activities. The proposed architecture consists of a Siamese transformer-based network that takes input pairs of videos and determines if they belong to the same activity or not. If they do, the videos are also temporally aligned. A key innovation of our approach is the direct action alignment between videos, crucial for accurately matching corresponding segments. This is enabled by the Siamese two-stage architecture that ensures robust initialization for temporal alignment. By explicitly modelling intra-video action discrimination, inter-video action associations, and inter-activity action associations, our method significantly outperforms state-of-the-art approaches on the challenging BF and YTI datasets.

Acknowledgements This work was supported by the grant PRE2020-094714, the project PID2019-110977GA-I00 and the project PID2023-151351NB-I00 funded by MCIN/ AEI /10.13039/501100011033, by "ESF Investing in your future" and by ERDF, UE.

References

1. Aakur, S.N., Sarkar, S.: A perceptual prediction framework for self supervised event segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
2. Alayrac, J.B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., Lacoste-Julien, S.: Unsupervised learning from narrated instruction videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4575–4583 (2016)
3. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: Deep Canonical Correlation Analysis (ICML). pp. 1247–1255 (2013)
4. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
5. Bahrami, E., Francesca, G., Gall, J.: How much temporal long-term context is needed for action segmentation? In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2023)
6. Behrmann, N., Golestaneh, S.A., Kolter, Z., Gall, J., Noroozi, M.: Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
7. Bueno-Benito, E., Tura, B., Dimiccoli, M.: Leveraging triplet loss for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
9. Chang, X., Tung, F., Mori, G.: Learning discriminative prototypes with dynamic time warping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

10. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. In: International conference on machine learning (ICML) (2017)
11. Dias, C., Dimiccoli, M.: Learning event representations by encoding the temporal context. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. vol. 11131, pp. 587–596 (2018)
12. Dimiccoli, M., Wendt, H.: Learning event representations for temporal segmentation of image sequences by dynamic graph embedding. IEEE Transactions on Image Processing **30**, 1476–1486 (2020)
13. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: An analysis of modern techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
14. Ding, G., Yao, A.: Temporal action segmentation with high-level complex activity labels. IEEE Transactions on Multimedia (2023)
15. Donahue, G., Elhamifar, E.: Learning to predict activity progress by self-supervised video alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
16. Du, Z., Wang, X., Zhou, G., Wang, Q.: Fast and unsupervised action boundary detection for action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
17. Dvornik, N., Hadji, I., Derpanis, K.G., Garg, A., Jepson, A.D.: Drop-dtw: Aligning common signal between sequences while dropping outliers. In: Advances in Neural Information Processing Systems (2021)
18. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
19. Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
20. Haresh, S., Kumar, S., Coskun, H., Syed, S.N., Konin, A., Zia, M.Z., Tran, Q.H.: Learning by aligning videos in time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
21. He, Y., Yuan, Z., Wu, Y., Cheng, L., Deng, D., Wu, Y.: Vistec: Video modeling for sports technique recognition and tactical analysis. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2024)
22. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
23. Kukleva, A., Kuehne, H., Sener, F., Gall, J.: Unsupervised learning of action classes with continuous temporal embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
24. Kumar, S., Haresh, S., Ahmed, A., Konin, A., Zia, M.Z., Tran, Q.H.: Unsupervised action segmentation by joint representation learning and online clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
25. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
26. Li, J., Todorovic, S.: Action shuffle alternating learning for unsupervised action segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

27. Li, S.J., AbuFarha, Y., Liu, Y., Cheng, M.M., Gall, J.: Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
28. Li, Y., Xue, Z., Xu, H.: Otas: Unsupervised boundary detection for object-centric temporal action segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)
29. Liu, W., Tekin, B., Coskun, H., Vineet, V., Fua, P., Pollefeys, M.: Learning to align sequential actions in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
30. Lu, Z., Elhamifar, E.: Weakly-supervised action segmentation and alignment via transcript-aware union-of-subspaces learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
31. Lu, Z., Elhamifar, E.: Set-supervised action learning in procedural task videos via pairwise order consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19871–19881 (2022)
32. Lu, Z., Elhamifar, E.: Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
33. Ng, Y.B., Fernando, B.: Weakly supervised action segmentation with effective use of attention and self-attention. Computer Vision and Image Understanding **213**, 103298 (2021)
34. Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., Schiele, B.: Recognizing fine-grained and composite activities using hand-centric features and script data. International Journal of Computer Vision (IJCV) **119**, 346–373 (2016)
35. Sarfraz, M.S., Murray, N., Sharma, V., Diba, A., Gool, L.V., Stiefelhagen, R.: Temporally-weighted hierarchical clustering for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
36. Sarfraz, M.S., Sharma, V., Stiefelhagen, R.: Efficient parameter-free clustering using first neighbor relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
37. Sener, F., Yao, A.: Unsupervised learning and segmentation of complex activities from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
38. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: IEEE International Conference on Robotics and Automation (ICRA) (2018)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
40. Souri, Y., Fayyaz, M., Minciullo, L., Francesca, G., Gall, J.: Fast weakly supervised action segmentation using mutual consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
41. Swetha, S., Kuehne, H., Rawat, Y.S., Shah, M.: Unsupervised discriminative embedding for sub-action learning in complex activities (2021)
42. Tran, Q.H., Mehmood, A., Ahmed, M., Naufil, M., Konin, A., Zia, M.Z.: Permutation-aware activity segmentation via unsupervised frame-to-segment alignment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)

43. VidalMata, R.G., Scheirer, W.J., Kukleva, A., Cox, D.D., Kuehne, H.: Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021)
44. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2013)
45. Wang, Z., Chen, H., Li, X., Liu, C., Xiong, Y., Tighe, J., Fowlkes, C.: Sscap: Self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2022)
46. Xu, A., Zheng, W.S.: Efficient and effective weakly-supervised action segmentation via action-transition-aware boundary alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
47. Xu, M., Gould, S.: Temporally consistent unbalanced optimal transport for unsupervised action segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
48. Yi, F., Wen, H., Jiang, T.: Asformer: Transformer for action segmentation. In: The British Machine Vision Conference (BMVC) (2021)
49. Zhang, R., Wang, S., Duan, Y., Tang, Y., Zhang, Y., Tan, Y.P.: Hoi-aware adaptive network for weakly-supervised action segmentation. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2023)
50. Zuckerman, I., Werner, N., Kouchly, J., Huston, E., DiMarco, S., DiMusto, P., Laufer, S.: Depth over rgb: automatic evaluation of open surgery skills using depth camera. International Journal of Computer Assisted Radiology and Surgery pp. 1–9 (2024)