Unsupervised Action Segmentation via Fast Learning of Semantically Consistent Actoms

Zheng Xing[1], **Weibing Zhao**[2]*

[1] Future Network of Intelligence Institute, School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

[2] Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China

{zhengxing, weibingzhao}@link.cuhk.edu.cn

Abstract

Action segmentation serves as a pivotal component in comprehending videos, encompassing the learning of a sequence of semantically consistent action units known as *actoms*. Conventional methodologies tend to require a significant consumption of time for both training and learning phases. This paper introduces an innovative unsupervised framework for action segmentation in video, characterized by its fast learning capability and absence of mandatory training. The core idea involves splitting the video into distinct actoms, which are then merging together based on shared actions. The key challenge here is to prevent the inadvertent creation of singular actoms that attempt to represent multiple actions during the splitting phase. Additionally, it is crucial to avoid situations where actoms associated with the same action are incorrectly grouped into multiple clusters during the merging phase. In this paper, we present a method for calculating the similarity between adjacent frames under a subspace assumption. Then, we employ a local minimum searching procedure, which effectively *splits* the video into coherent actoms aligned with their semantic meaning and provides us an action segmentation proposal. Subsequently, we calculate a spatio-temporal similarity between actoms, followed by developing a merging process to *merge* actoms representing identical actions within the action segmentation proposals. Our approach is evaluated on four benchmark datasets, and the results demonstrate that our method achieves state-of-the-art performance. Besides, our method also achieves the optimal balance between accuracy and learning time when compared to existing unsupervised techniques. **Code** is available at *https://github.com/y66y/SaM*.

Introduction

Large volumes of videos are uploaded to both cloud and edge storage every day, leading to a significant demand for rapid video analysis. Efficient video comprehension plays a pivotal role in real-world applications, such as video retrieval, surveillance analysis (Vishwakarma and Agrawal 2013), robot perception (Qi et al. 2019; **?**, 2021; Sun et al. 2023, 2022b,c,a), indoor localization (Wang et al. 2021, 2020a; Liu, Wang, and Luo 2020; Luo, Zhang, and Wang 2020). In recent years, a considerable focus within the field of video comprehension has been devoted to action segmentation in videos (Wang et al. 2023; Sheng and Li 2023; van Amsterdam et al. 2023). The objective of action segmentation involves categorizing concise, pre-edited segments that characterize individual actions. Despite significant progress in supervised action segmentation techniques, driven by the emergence of deep neural networks and extensive datasets, models based on fully supervised learning still require laborious manual data annotation. This process is time-consuming, costly, and susceptible to errors. Consequently, unsupervised action segmentation has emerged as an alternative strategy to tackle this challenge.

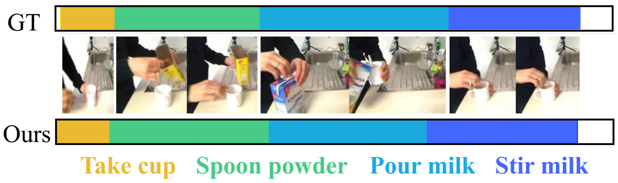


Figure 1: Action segmentation output example from Breakfast Dataset (Kuehne, Arslan, and Serre 2014): *P46_webcam02_P46_milk*. Colors indicate different actions in chronological order: take_cup, spoon_powder, pour_milk, stir_milk. The background is shown in white color.

Action segmentation involves assigning action labels to individual frames within a video sequence, typically depicting a person engaging in a series of actions as part of a higher-level activity. An illustrative example of breakfast preparation is depicted in Fig. 1. Compared to recognizing activities in videos, action segmentation introduces more formidable challenges due to the presence of extraneous background frames. A significant obstacle arises from the necessity for a substantially larger number of annotations to effectively guide learning-based methodologies, which has resulted in the popularity of weakly supervised and unsupervised approaches for action segmentation (Wang et al. 2023; de AP Marques et al. 2022; Sheng and Li 2023; Li, He, and Xu 2022). Some techniques utilize textual information extracted from accompanying audio to assign action labels at the frame level for training action segmentation models, as introduced in (Alayrac et al. 2016). However, this approach relies on the assumption of well-synchronized

*Corresponding author.

audio and video frames. Alternatively, other methodologies presuppose some *prior knowledge* of actions, such as the high-level activity labels or lists of depicted actions in each video (Souri et al. 2021). However, even this level of annotation demands substantial annotation effort for each training video, due to the variability of constituent actions across diverse activities. Regardless of the level of prior knowledge, the majority of weakly- and unsupervised methods, concentrate on obtaining pseudo-labels, which subsequently supervise the training of task-specific feature embeddings. However, the acquired pseudo-labels are inherent noisy, which may potentially impede the effectiveness of the learned embeddings

This paper introduces an innovative unsupervised action segmentation framework comprising two distinct phases: splitting and merging. Our approach is grounded in two fundamental insights. Firstly, we assume that high-dimensional frames within action videos reside in distinct subspaces, each corresponding to a specific action. Secondly, humans typically perceive frame segments as manifestations of individual actions. Building upon this comprehension, the identification of actoms arises as an effective and efficient approach for segmenting actions within lengthy, untrimmed videos. Leveraging these insights, our algorithm begins by dividing the video into multiple actoms, where each actom encapsulates frames representing a particular action. Nonetheless, considering the potential recurrence of the same action multiple times within a single video (for instance, a dance video featuring actions A and B followed by another instance of action A), there is a necessity to merge actoms associated with identical actions.

The primary challenge lies in effectively preventing a single actom from erroneously encapsulating multiple actions during the splitting phase. It is also equally important to avoid the situation where actoms linked with the same action are incorrectly grouped into separate actions during the subsequent merging process. Specifically, our objective is to ensure that each distinct actom obtained through the video-splitting process accurately encapsulates only one specific action. Given the intricate dynamics of motion backgrounds and the inherent variations in action execution, the task of splitting videos containing *non-clustered* frames into discrete actoms poses a significant challenge. Furthermore, during the actom merging phase, it is challenging to prevent the fusion of two actoms representing different actions. Our strategy relies on capturing the intricate relationship between these actoms with precision. However, accurately evaluating the degree of similarity between these two actoms is a complex task.

In this paper, we draw inspiration from the Canny detector (Canny 1986), commonly used in image processing, and the segmentation method (Keogh et al. 2004), applied in time series analysis. Our initial effort involves identifying distinctive features that exhibit coherence within a specific action context while manifesting variability when compared against different actions. However, it's essential to acknowledge that challenges like occlusion, shifts in viewpoints, or fluctuations in lighting can result in temporal features derived from actions lacking strict uniformity. Diverging from the methodology of the Canny detector, which identifies intensity gradients within images, we obtain an understanding of actions within videos by learning the actom in the video. Specifically, we identify potential boundaries of actoms through a comprehensive evaluation of the subspace-based similarity between consecutive frames. One of the previously mentioned challenges is to prevent a single actom from erroneously encapsulating multiple actions during the splitting phase. To tackle this challenge, we propose to utilize the minimum value selected from localized temporal windows on the similarity curve to establish the boundaries of actoms. These identified boundaries will then serve as guides for segmenting the video into distinct, semantically consistent actoms. During the actom merging phase, a challenge arises in ensuring that actoms associated with the same action are not incorrectly separated into distinct actions. Accurately quantifying the degree of similarity between two such actoms is crucial. Therefore, we introduce a novel spatio-temporal similarity measure between actoms, considering both their temporal separations and appearance feature distances and facilitating the fast amalgamation of actoms into cohesive actions. Our work contains the following main contributions:

- We introduce a novel unsupervised learning framework for action segmentation, consisting of two essential components: a splitting procedure and a merging procedure. The splitting procedure ensures the precise division of the video into distinct actoms, while the merging procedure guarantees the aggregation of actoms that represent the same action into coherent clusters.
- First, compared to traditional unsupervised methodologies that heavily rely on pseudo-labels for supervised training, our approach distinguishes itself by completely bypassing the necessity for any form of training, thus possessing the advantage of speed. Second, our method adeptly leverages the semantically consistent attributes of the temporal frames within action videos. Specifically, we employ a splitting procedure to partition the video into concise actoms. This process ensures that each actom comprises frames with similar semantic characteristics. Under the premise that actoms in close temporal proximity are more likely to exhibit similar semantic traits, we propose merging these actoms based on an effective spatio-temporal similarity measure between them.
- Our proposed method skillfully achieves a balanced trade-off between model accuracy and learning speed, outperforming weakly supervised and unsupervised action segmentation techniques across four benchmark datasets. Remarkably, our method demonstrates comparable performance even when compared to supervised methods.

Related Work

Action Segmentation in videos has attracted significant research interest, as evidenced by the considerable volume of related studies (Bueno-Benito, Vecino, and Dimiccoli

2023; van Amsterdam et al. 2023). In this section, our attention is directed towards a comprehensive review of existing methodologies relevant to the challenge of action segmentation. We particularly emphasize approaches about weakly-supervised and unsupervised paradigms.

Existing methods for action segmentation can be broadly categorized into three groups: fully supervised (Liu et al. 2022; van Amsterdam et al. 2023; Lim et al. 2023), weakly supervised (Souri et al. 2021; Sheng and Li 2023; Luo et al. 2022; Fayyaz and Gall 2020), and unsupervised (Kukleva et al. 2019; Bueno-Benito, Vecino, and Dimiccoli 2023; Wang et al. 2023). They differ in whether the annotations are collected by human annotators or extracted in a semi- or unsupervised manner. These models typically follow a paradigm where an embedding is trained on top of pre-extracted frame-level video features, as seen in the works of Sheng et al. (Sheng and Li 2023), Sener et al. (Sener and Yao 2020), and Richard et al. (Richard et al. 2018), or hand-crafted video features, as demonstrated by Ding et al. (Ding and Xu 2018) and Kukleva et al. (Kukleva et al. 2019). The training process of the embedding layer involves the use of a discriminative objective function in conjunction with available annotations (Li, Lei, and Todorovic 2019; Sheng and Li 2023; Richard et al. 2018; Souri et al. 2021). In the subsequent sections, we explore the specifics of weakly supervised and unsupervised techniques, which differ significantly in how they extract and exploit pseudo-labels.

Weakly-supervised approaches often assume the availability of both the activity label at the video level and action ordering, referred to as transcripts, during the training phase. Many methods follow a two-step procedure: initially generating pseudo-labels using transcripts and subsequently training a frame classification network with these inferred labels (Sheng and Li 2023; Luo et al. 2022). In contrast, NN-Vit (Richard et al. 2018) directly utilizes transcripts to train a frame classification model. To enforce consistency between frame-level label predictions, they introduce a loss based on Viterbi decoding. In a similar vein, MuCoN (Souri et al. 2021) aims to leverage transcripts in learning a frame classification model. They employ two network branches, with only one having access to transcripts, ensuring mutual consistency between both branches. Another recent method, CDFL (Li, Lei, and Todorovic 2019), also seeks to utilize transcripts in training its frame labeling model. Initially, they construct a fully-connected, directed segmentation graph, where paths represent actions. Training the model involves maximizing the energy difference between valid paths (i.e., paths consistent with the ground-truth transcript) and invalid ones. In SCT (Fayyaz and Gall 2020), the authors assume knowledge of the set of action labels for a given video, but without their order. They determine the ordering and temporal boundaries of actions by alternately optimizing set and frame classification objectives. This ensures that frame-level action predictions align with set-level predictions.

Unsupervised approaches typically rely solely on knowledge of the video-level activity label (Bueno-Benito, Vecino, and Dimiccoli 2023; VidalMata et al. 2021; Aakur and Sarkar 2019). The Mallow method (Sener and Yao 2020) utilizes video-level annotations in an iterative approach to action segmentation. This involves alternating optimization between a discriminative appearance model and a generative temporal model of action sequences. Conversely, the Frank-Wolfe (Alayrac et al. 2016) method extracts video narrations using Automatic Speech Recognition (ASR). These narrations are then employed to extract an action sequence for a set of videos related to a specific activity. This is achieved by independently clustering the videos and the ASR-recovered speech to identify action verbs in each video. Temporal localization is subsequently obtained by training a linear classifier. CTE proposes learning frame embeddings that incorporate relative temporal information. They train a video activity model using pseudo-labels generated from K-means clustering of the videos' IDT features. The trained embeddings are then re-clustered to match the groundtruth number of actions, and their order is determined using statistics of the relative timestamps with GMM+Viterbi decoding. VTE-UNET (VidalMata et al. 2021) leverages similarly learned embeddings, combining them with temporal embeddings to enhance the performance of CTE. LSTM+AL (Aakur and Sarkar 2019) fine-tunes a pre-trained VGG16 model with an LSTM, using future frame prediction as a self-supervision objective to learn frame embeddings. These embeddings are subsequently employed to train an action boundary detection model.

However, all these methods necessitate training on the target video dataset, which, from a practical standpoint, imposes significant restrictions. In contrast, our method eliminates the need for training and relies solely on video splitting and merging.

Methodology

Observed that with a well-established similarity between frames in the video, the boundaries of actoms within a video can be identified without resorting to additional training on objectives reliant on objectives that use noisy pseudo-labels, something that almost all current methods pursue. Previous endeavors in actom boundary detection have involved complex neural networks or the generation of pseudo-labels that may not be directly relevant (Ishikawa et al. 2021; Wang et al. 2020b). Contrary to this prevailing trend, our approach takes a different path.

The bottom-up framework, exemplified by (Menon, Muthukrishnan, and Kalyani 2020), emerges as a promising choice for our task. Such methods furnish a hierarchy of data partitions instead of a singular partition. In this paper, we embrace a bottom-up framework for action segmentation, bypassing the need for video-level activity labels. The capacity of our approach to generating plausible action segmentation without training holds considerable practical value.

Given a video $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, our primary goal is to categorize these frames into K actions, where K represents the number of distinct actions present in the video. Our approach begins with an in-depth explanation of frame similarity. Subsequently, we outline the methodology for detecting actom boundaries based on the similarity between adjacent frames, a process that divides videos into discrete and semantically consistent actoms. Moreover, we assume that actoms located closely in the temporal domain share similar semantic characteristics. To facilitate this, we introduce a spatio-temporal similarity measure that forges connections between actoms, taking into account their proximity not only in feature space but also in the temporal dimension. This fusion of feature-based proximity and temporal alignment aims to incorporate spatial and temporal coherence in a unified manner.

Splitting Video into Semantically Consistent Actoms

Our initial objective revolves around partitioning a video into actoms by accurately identifying their boundaries. However, due to factors such as occlusion, changes in viewpoint, or variations in lighting, the alterations along the entire temporal dimension of a video can be abrupt. Consequently, precisely detecting actom boundaries based solely on the general distance assessment (e.g., Euclidean) between consecutive frames poses considerable challenges.

The subspace assumption has found extensive application in various domains, including image representation and compression, as well as in addressing computer vision challenges like action segmentation, face clustering, image segmentation, and video segmentation. Our work aligns with these principles and similarly operates under the premise that distinct actions present in video frames can be discerned by their respective placements within distinct subspaces. We introduce a cosine-based measurement to quantify the similarity between frames, building upon the subspace assumption. Specifically, we measure the similarity between frames within the same or different subspaces by evaluating the cosine angle between them. To achieve this, we first normalize each frame $\mathbf{x}_i$ as $\tilde{\mathbf{x}}_i = \mathbf{x}_i/||\mathbf{x}_i||^2$, where $||\cdot||$ denotes the l_2 norm. These normalized frames $\tilde{\mathbf{x}}_i \in \mathbb{R}^D$, with $i \in 1, 2, ..., N$, are situated within the high-dimensional hypersphere $\mathbb{S}^{D-1}$ (Menon, Muthukrishnan, and Kalyani 2020). We define the angle $\theta_{i,j}$ between two frames $\mathbf{x}_i$ and $\mathbf{x}_j$ as $\theta_{i,j} = \cos^{-1}(\tilde{\mathbf{x}}_i^{\mathrm{T}}\tilde{\mathbf{x}}_j)$, where $\theta_{i,j} \in [0, \pi]$. This radian-based $\theta_{i,j}$ is then converted to degrees, denoted as $\tilde{\theta}_{i,j}$, using the formula $\tilde{\theta}_{i,j} = \theta_{i,j} \cdot 180/\pi$, with $\tilde{\theta}_{i,j} \in [0, 180]$. The angle $\tilde{\theta}_{i,j} = 0$ when frames $\mathbf{x}_i$ and $\mathbf{x}_j$ reside within the same subspace, while $\tilde{\theta}_{i,j} > 0$ indicates that $\mathbf{x}_i$ and $\mathbf{x}_j$ are located in different subspaces. The similarity between consecutive frames $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ is defined as $s_i = \exp(-\tilde{\theta}_{i,i+1}/\sigma_\theta^2)$ for $i \in [1, N-1]$ and s_N is set to be $s_N = s_{N-1}$ for convenience. In this paper, the variance σ_θ^2 is designated as $\sigma_\theta^2 = \mathrm{Var}[\tilde{\theta}]$.

If we graph the similarity s_i against their corresponding time stamps i, we would observe a sequence of undulating patterns resembling the shape of $\sqcap$. The fluctuations in similarity over time can be attributed to the fact that actions belonging to the same category exhibit high similarity values (approaching 1), while significant decreases in similarity values indicate substantial changes in actions. Consequently, the boundaries represented by the falling edge of $\sqcap$ are the critical delineations between different actoms. We can simply locate the low value in the curve of s_i to identify these boundaries of actoms. However, in practice, pinpointing these troughs on the curve of s_i is not straightforward due to various factors. In particular, the regions around the boundaries of actions tend to be plagued by a multitude of erroneous responses. To address this issue, we develop a local minimum search algorithm to mitigate the influence of these errors. Specifically, we embark on a comprehensive boundary search by constructing a set of boundaries as

$$\{ \underset{t\in\{i+1,i+2,...,i+L\}}{\operatorname{argmin}} s_t | i \in \{0, L, 2L, 3L, ...N-L\}\}$$

where the window size $L = \lfloor \delta N/K \rfloor$, the length of the data N, and the number of actions K. In this construction, we search for minima within locally prominent windows of size L along the temporal dimension. These identified minima hold the potential to serve as effective boundaries between adjacent actoms. In the upcoming experimental section, we will conduct a detailed analysis of how the window size δ affects the performance of our algorithm. Furthermore, we will illustrate the robustness of our algorithm in response to changes in δ.

Algorithm 1: The split-and-merge (SaM) algorithm.

```
Input: the video $\mathcal{X}$, and the number of actions $K$.
Output: the $K$ actoms.
Calculate the similarity between adjacent frames: $\{s_i\}$.
Local minimum searching resulting in the segmentation indices $\mathcal{B} = \{b_1, b_2, ..., b_{M-1}\}$.
Initialize the actom $\{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_M\}$ with the corresponding index set $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_M\}$ according to $\mathcal{B}$.
repeat
    Compute the actom feature $\{\bar{\mathbf{x}}_m\}_{m\in[1,2,..,M]}$, and averaged time-stamps $\bar{t}_m = \frac{1}{|\mathcal{C}_m|}\sum_{t\in\mathcal{C}_m} t$.
    Compute the spatio-temporal similarity $G^{(M)}(i, j)$ for any $i \neq j$.
    Merge the most similar $i$th, and $j$th actoms, resulting in new actom.
until the number of actoms is $K$.
```

Merging Actoms Representing the Same Action

Denote the boundary set obtained from the previous section as $\mathcal{B} = \{b_1, b_2, \ldots, b_{M-1}\} \subset \{1, 2, \ldots, N\}$. We assume that the boundaries are both ordered and non-repeating, such that $b_{k-1} < b_k$ for any $k \in [1, M]$. The dummy boundaries are implicitly available: $b_0 := 0$ and $b_M := N$. Since the same action usually occurs multiple times in a video, M is always greater than the number of actions K. Therefore, it is necessary to develop a merging procedure to further group actoms into K actions.

Denote the mth actom as $\mathcal{X}_m = \{\mathbf{x}_i\}_{i\in\mathcal{C}_m}$, where the index set $\mathcal{C}_m = \{b_{m-1}+1, b_{m-1}+2, ..., b_m\}$. The key to merging actoms lies in measuring the similarity between them. Drawing upon the observation of identifying linking chains within data through the presence of nearest or shared neighbors, we introduce a spatio-temporal measurement that considers both proximities in the feature space and the temporal arrangement of actoms. Specifically, we aim to design a similarity metric that captures the essence of both feature-space and temporal closeness among actoms. This is achieved by incorporating the progression of time as a modulating factor during the similarity computation. We first compute the feature vector of actoms, denoted as $\{\bar{\mathbf{x}}_m\}_{m\in[1,2,..,M]}$, where $\bar{\mathbf{x}}_m = \frac{1}{|\mathcal{C}_m|}\sum_{i\in\mathcal{C}_m}\mathbf{x}_i$ represents the feature of the mth actom that characterizes a specific action. The similarity between the ith and jth actoms is then defined as

$$G^{(M)}(i,j) = \exp(-\frac{\lambda|\bar{t}_i - \bar{t}_j|}{N}) \cdot \exp[-\frac{180}{\pi\sigma_\theta^2}\cos^{-1}(\frac{\bar{\mathbf{x}}_i^{\mathrm{T}}\bar{\mathbf{x}}_j}{||\bar{\mathbf{x}}_i||^2||\bar{\mathbf{x}}_j||^2})]$$

for any $i, j \in [1, 2, ..., M]$, $i \neq j$, where $\bar{t}_m = \frac{1}{|\mathcal{C}_m|}\sum_{t\in\mathcal{C}_m} t$, and λ serves as a trade-off parameter, offering control over the influence of the temporal consistency requirement.

The introduced term $\exp(-\frac{\lambda|\bar{t}_i - \bar{t}_j|}{N})$ is employed to accentuate the significance of the temporal consistency on the similarity measure. Here, $|\bar{t}_i - \bar{t}_j|$ represents the temporal disparity between the respective nodes, with its impact adjusted by λ. This factor is especially pertinent as we intend to employ temporal similarity as a modulating component for the feature-space similarity. Consequently, $G^{(M)}(i,j)$ signifies the spatio-temporal similarity between the actoms i and j, taking into account the temporal relationships while considering the weight derived from the duration of the video sequence.

The entry $G^{(M)}(i,j)$, for any $i, j \in \{1, 2, ..., M\}$, $i \neq j$ with the maximum value is selected and the corresponding $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}_j$ are considered as the most similar actoms, which are merged to generate a new actom $\mathcal{X}_{m'} = \{\mathbf{x}_i\}_{i\in\mathcal{C}_{m'}}$, where the index set $\mathcal{C}_{m'} = \mathcal{C}_i \cup \mathcal{C}_j$. We repeat the merging process until the number of actoms reduces to K. The main steps of the proposed algorithm are shown in Alg. 1.

Experiments

Experimental Setup

Datasets. We assessed the efficacy of our approach on four benchmark datasets: Breakfast (BF)(Kuehne, Arslan, and Serre 2014), YouTube Instructional Videos (YTI) (Alayrac et al. 2016), Hollywood Extended (HE) (Bojanowski et al. 2014), and 50Salads (FS) (Stein and McKenna 2013). These four datasets encompass a broad spectrum of activities, ranging from diverse cooking routines to tasks like car maintenance. The dataset characteristics span varying video lengths, with averages ranging from approximately 520 frames to as high as 11788 frames.

Features. To ensure an equitable comparison with relevant prior studies, we adopt the same input features as recent methodologies (Sheng and Li 2023; Wang et al. 2023; Fayyaz and Gall 2020). Specifically, for the BF, FS, and HE datasets, we utilize the Improved Dense Trajectory (IDT) features (Wang and Schmid 2013) as computed and provided by the authors of CTE (Kukleva et al. 2019) (for BF and FS) and SCT (Fayyaz and Gall 2020) (for HE). For YTI (Alayrac et al. 2016), we leverage the features made available by the authors themselves. These features consist of 3000-dimensional vectors, achieved by concatenating Histogram of Optical Flow (HOF) (Laptev et al. 2008) descriptors with feature embeddings extracted from VGG16-conv5 (Simonyan and Zisserman 2014). Throughout all datasets, our reporting of performance encompasses the entire dataset, ensuring alignment with established practices within the literature.

Evaluation Metrics. Since our method outputs clusters without particular correspondences to the ground-truth labels, we require a one-to-one mapping between the outputs and the ground-truth labels. Following (Aakur and Sarkar 2019; Kukleva et al. 2019; Sener and Yao 2020), we utilize the Hungarian algorithm to generate this mapping based on the overlap between matched clusters. Since our method does not concern cluster labels, we conduct this mapping on the video level as in (Aakur and Sarkar 2019). We also report the F1 score and mean over frames (MoF) for all datasets as used in previous works (Kukleva et al. 2019). We report the Jaccard index as an intersection over union (IoU) as an additional measurement.

Comparison to State-of-the-art

We proceed to present a comprehensive comparison of our method against the current state-of-the-art techniques, including WPI (Ghoddoosian et al. 2022), SSTDA (Chen et al. 2020), ASAL (Li and Todorovic 2021), TOT+TCL (Kumar et al. 2022), Mallow (Sener and Yao 2020), ASAL (Li and Todorovic 2021), SCV (Li and Todorovic 2020), US-FGW (Luo et al. 2022), DMR (Asghari-Esfeden, Sznaier, and Camps 2020), D3TW (Chang et al. 2023), SRL (Feichtenhofer et al. 2021), SRA (Lai et al. 2019), STPE (de AP Marques et al. 2022), GMM+CNN (Kuehne, Richard, and Gall 2019), FFA (Ng and Fernando 2020), C2F (Sheng, Li, and Tian 2021), TAD (Li, He, and Xu 2022), etc. We will discuss the results individually for each of the four datasets, as summarized in the following tables: Tab. 1 (BF), Tab. 2 (YTI), Tab. 4 (FS), and Tab. 3 (HE). However, it's important to acknowledge that as highlighted in (Kukleva et al. 2019), while our evaluation metrics are comparable to those utilized by both weakly and fully supervised approaches, a certain nuance must be taken into account. Specifically, the results of unsupervised learning are presented concerning an optimal cluster assignment to ground-truth classes, thereby representing the best conceivable scenario for this task. For each dataset, we incorporate partial relevant metrics whenever they are conventionally utilized for that specific dataset. In the presented tables, the *Train* column serves as an indicator of whether the method necessitates training on the target activity videos before executing the segmentation process. A hyphen – denotes instances where no reported results are available.

Performance on BF. In Tab. 1, we present a performance comparison with state-of-the-art methods on BF. In addition to unsupervised methods, we also include a comparison with several supervised and weakly supervised methods, which serve as upper bounds for evaluating our method's performance. Our SaM method demonstrates superior performance over all unsupervised methods, showcasing absolute improvements of 9.5%, 10.2%, and 12% compared to the best-reported unsupervised method CoSeg (Wang

BF				
Weakly Supervised	**IoU**	**F1**	**MoF**	**Train**
CDFL	–	–	50.2	✓
SCT	–	–	30.4	✓
MuCon	–	–	48.5	✓
WPI	25.0	–	–	✓
C2FL	–	–	50.4	✓
Unsupervised	**IoU**	**F1**	**MoF**	**Train**
CTE	–	26.4	41.8	✓
SSTDA	–	–	55.2	✓
ASAL	–	37.9	52.5	✓
TOT+TCL	–	30.3	39.0	✓
CoSeg	42.6	44.7	53.1	✓
SaM	**44.4**	**55.9**	**64.0**	×

Table 1: Comparison to the state-of-the-art on BF.

YTI			
Unsupervised	**F1**	**MoF**	**Train**
CTE	28.3	39.0	✓
Mallow	27.0	27.8	✓
ASAL	32.1	44.9	✓
TOT+TCL	32.9	45.3	✓
LTL	34.7	52.4	×
SaM	**49.6**	**68.1**	×

Table 2: Comparison to the state-of-the-art learning on YTI.

et al. 2023). Similarly, our approach outperforms the leading weakly supervised method C2FL (Sheng and Li 2023) with a remarkable 14.7% enhancement on the MoF metric. Furthermore, in comparison with fully supervised methods, our approach achieves results 5.9% lower than the F1 metric of (van Amsterdam et al. 2023) and 5.1% lower than the MoF metric of (Chen et al. 2020). Although our method falls short of the performance exhibited by fully supervised methods, we have achieved remarkable proximity to their results.

Performance on YTI. We summarize the performance of our method on YTI in Tab. 2. To make a fair comparison, we remove background frames from videos as previous approaches (Kukleva et al. 2019; Sener and Yao 2020). Our SaM method significantly outperforms all unsupervised methods, with absolute gains of 14.9%/15.7% on F1/MoF over the best published unsupervised method LTL (Bueno-Benito, Vecino, and Dimiccoli 2023).

Performance on HE. As shown in Tab. 3, our method achieves a significant leap compared with existing approaches. In particular, our method obtains improvements of 5.7% on MoF compared with the best unsupervised method DHC (Sharma et al. 2023). Similarly, our method outperforms the best weakly supervised method MuCon (Souri et al. 2021) with 19.3% gains on the MoF metric. Remarkably, our method even outperforms all fully supervised methods on the IoU and MoF metrics.

HE				
Fully Supervised	**IoU**	**F1**	**MoF**	**Train**
GMM+CNN	8.4	–	39.5	✓
SCV	35.5	–	–	✓
Weakly Supervised	**IoU**	**F1**	**MoF**	**Train**
CDFL	19.5	–	40.6	✓
SCT	17.7	–	–	✓
MuCon	–	–	41.6	✓
US-FGW	23.1	–	38.4	✓
D3TW	–	–	33.6	✓
Unsupervised	**IoU**	**F1**	**MoF**	**Train**
SRA	–	41.1	–	✓
DMR	–	47.4	–	✓
SRL	30.0	45.7	53.0	✓
STPE	–	–	47.3	✓
DHC	–	–	55.2	✓
SaM	**39.4**	**57.3**	**60.9**	×

Table 3: Comparison to the state-of-the-art on HE.

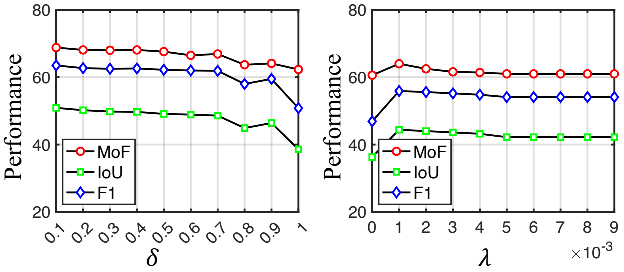
FS				
Weakly Sup.	**F1**	**MoF**(eval)	**MoF**(mid)	**Train**
CDFL	–	–	54.7	✓
FFA	–	–	49.4	✓
C2F	–	–	24.7	✓
TAD	–	–	45.5	✓
C2F2	–	–	56.2	✓
Unsup.	**F1**	**MoF**(eval)	**MoF**(mid)	**Train**
CTE	–	35.5	30.2	✓
SSTDA	73.8	–	–	✓
ASAL	–	39.2	34.4	✓
TOT+TCL	–	44.5	34.3	✓
CoSeg	71.8	–	–	✓
SaM	**78.2**	**71.6**	**71.9**	×

Table 4: Comparison to the state-of-the-art on FS.

Performance on FS. We provide a summary of the performance of our method on the FS dataset in Tab. 4. The FS dataset encompasses an average of 19 actions per video, with 14.1% of all frames classified as background frames. We evaluate our method based on two levels of action granularity, as outlined in (Stein and McKenna 2013). The *mid* granularity level assesses performance across the complete set of 19 actions, while the *eval* granularity level combines certain action classes to yield 10 distinct action classes. In the *mid* granularity evaluation, our method achieves an MoF of 71.9%, which is notably higher by 15.7% (in absolute terms) compared to the leading weakly supervised method C2F2 (Sheng and Li 2023). This trend of performance improvement is also evident in the *eval* granularity level evaluation. In comparison to the most effective fully supervised method, our approach achieves results that are 10.29% lower in terms of F1 metric, 12.2% lower in terms of *eval* metric, and 4.4% lower in terms of *mid* metric. This indicates that

	Train (h)	**Learn** (s)	**Train**
Weakly Supervised			
CDFL	66.73	62.37	✓
MuCon-full	4.57	3.03	✓
Unsupervised			
CTE	–	217.94	✓
Ours	**0.00**	**0.20**	×

Table 5: Comparison of training and learning time. The training time is measured for training on split 1 on BF and the learning time is measured as the average learning time for a single video.



(a) (b)

Figure 2: The performance of our method with varying (a) δ and (b) λ on BF dataset.

there is still room for enhancement in our method's performance on the FS dataset.

Run-Time Comparison

The comparison of runtime efficiency between our method and alternative approaches is presented in Tab. 5, including MuCon-full (Souri et al. 2021), etc. All experiments were conducted using split 1 of the BF dataset, with learning times reported for individual videos. Each video within this dataset comprises approximately 2,000 frames. The time used for feature extraction is not included. A discernible pattern emerges from the table: our method eliminates the necessity for hours of GPU-intensive model training. In comparison to the faster unsupervised approach CTE (Kukleva et al. 2019), which either requires no training or entails training overhead, our method achieves an impressive 56x faster learning rate. The combination of swift learning speed and the absence of training prerequisites underscores the practical feasibility of our approach. When paired with an off-the-shelf feature extractor, our method becomes readily applicable to real-world applications.

Parameter Sensitivity Analysis

Impact of δ. We delve into investigating the impact of the window size parameter, δ, which governs the search for boundaries of actoms. We fix the parameter λ as $\lambda = 0.001$, The relationship between action segmentation performance and the parameter δ is illustrated in Fig. 2 (a). When δ is increased, our method tends to yield stable results, particularly within the range of $\delta < 0.4$. However, when δ exceeds 0.4, the effectiveness of our algorithm begins to diminish. This decline can be attributed to the fact that with larger δ values, the algorithm detects fewer boundaries of actoms. Nonetheless, it's worth noting that our algorithm maintains a relatively stable performance across a wide range of δ values, $0.1 < \delta < 0.4$. Our algorithm exhibits consistent behavior within this range. We set $\delta = 0.3$ in our experiment.

Influence of λ. We further delve into investigating the impact of the trade-off parameter, λ, on the actom merging process. We maintain a fixed value for the parameter δ, setting it to $\delta = 0.3$. The correlation between action segmentation performance and the parameter λ is depicted in Fig. 2 (b). Similar to δ, it appears that the parameter λ also exerts a significant influence on the performance of our algorithm. We notice that the performance of our algorithm improves as λ increases, and this trend starts to diminish when λ surpasses 0.001. However, there is a potential issue to consider. Increasing the value of λ to enhance temporal consistency can inadvertently lead to the merging of adjacent atoms that actually represent distinct actions. For example, if a video showcases actions A, B, and C, and the sequential occurrence of these actions follows the pattern A, B, A, C, B. In this case, a tremendous value of λ would result in merging the first three actoms A, B, and A into an action. To address this concern, it is advisable to adopt a balanced approach and refrain from assigning excessively large values to λ. A commonly suggested value for λ is 0.001.

Ablation Study

Impact of Splitting. We demonstrate the role of the splitting step by removing it from our algorithm. We assume that each frame represents an actom, and subsequently, our merging method is applied to these actoms. This adaptation results in a performance decline in our method on the BF dataset, with the MoF score dropping from 64.0 to 54.8, IoU decreasing from 44.4 to 35.1, and F1 diminishing from 55.9 to 39.6. This illustrates that our method cannot function effectively without the splitting step.

Impact of Merging. To illustrate the impact of the merging step, we exclude it from our algorithm. We execute the boundary search process, retaining only $K - 1$ boundaries with the lowest similarity s_i to ensure the generation of K segments, with each segment corresponding to an action. This adjustment leads to a performance decline in our method on the BF dataset, with the MoF score dropping from 64.0 to 50.6, IoU decreasing from 44.4 to 22.0, and F1 diminishing from 55.9 to 30.4. This emphasizes that our method cannot operate effectively without the merging step.

Conclusion

This paper introduces an innovative unsupervised learning framework for action segmentation. By considering both the consistency within actions and the variation across actions, we develop a SaM algorithm to learn the semantically consistent actoms. Rigorous evaluations conducted on four demanding datasets substantiate the efficacy of our approach.

References

Aakur, S. N.; and Sarkar, S. 2019. A perceptual prediction framework for self supervised event segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1197–1206.

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

Asghari-Esfeden, S.; Sznaier, M.; and Camps, O. 2020. Dynamic motion representation for human action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 557–566.

Bojanowski, P.; Lajugie, R.; Bach, F.; Laptev, I.; Ponce, J.; Schmid, C.; and Sivic, J. 2014. Weakly supervised action labeling in videos under ordering constraints. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 628–643.

Bueno-Benito, E. B.; Vecino, B. T.; and Dimiccoli, M. 2023. Leveraging triplet loss for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4921–4929.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

Chang, C.-Y.; Huang, D.-A.; Sui, Y.; Fei-Fei, L.; and Niebles, J. C. 2023. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3546–3555.

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

de AP Marques, G.; Busson, A. J. G.; Guedes, A. L. V.; Duarte, J. C.; and Colcher, S. 2022. Unsupervised method for video action segmentation through spatio-temporal and positional-encoded embeddings. In *Proceedings of the 13th ACM Multimedia Systems Conference*, 136–149.

Ding, L.; and Xu, C. 2018. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6508–6516.

Fayyaz, M.; and Gall, J. 2020. Sct: Set constrained temporal transformer for set supervised action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 501–510.

Feichtenhofer, C.; Fan, H.; Xiong, B.; Girshick, R.; and He, K. 2021. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3299–3309.

Ghoddoosian, R.; Dwivedi, I.; Agarwal, N.; Choi, C.; and Dariush, B. 2022. Weakly-supervised online action segmentation in multi-view instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13780–13790.

Ishikawa, Y.; Kasai, S.; Aoki, Y.; and Kataoka, H. 2021. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2322–2331.

Keogh, E.; Chu, S.; Hart, D.; and Pazzani, M. 2004. Segmenting time series: A survey and novel approach. In *Data mining in time series databases*, 1–21.

Kuehne, H.; Arslan, A.; and Serre, T. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 780–787.

Kuehne, H.; Richard, A.; and Gall, J. 2019. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding*, 163: 78–89.

Kukleva, A.; Kuehne, H.; Sener, F.; and Gall, J. 2019. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12066–12074.

Kumar, S.; Haresh, S.; Ahmed, A.; Konin, A.; Zia, M. Z.; and Tran, Q.-H. 2022. Unsupervised action segmentation by joint representation learning and online clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20174–20185.

Lai, Q.; Wang, W.; Sun, H.; and Shen, J. 2019. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE Transactions on Image Processing*, 29: 1113–1126.

Laptev, I.; Marszalek, M.; Schmid, C.; and Rozenfeld, B. 2008. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Li, J.; Lei, P.; and Todorovic, S. 2019. Weakly supervised energy-based learning for action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6243–6251.

Li, J.; and Todorovic, S. 2020. Set-constrained viterbi for set-supervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10820–10829.

Li, J.; and Todorovic, S. 2021. Action shuffle alternating learning for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12628–12636.

Li, Z.; He, L.; and Xu, H. 2022. Weakly-Supervised Temporal Action Detection for Fine-Grained Videos with Hierarchical Atomic Actions. In *European Conference on Computer Vision*, 567–584.

Lim, K. M.; Lee, C. P.; Tan, K. S.; Alqahtani, A.; and Ali, M. 2023. Fine-Tuned Temporal Dense Sampling with 1D Convolutional Neural Network for Human Action Recognition. *Sensors*, 23(11): 5276.

Liu, C.; Wang, C.; and Luo, J. 2020. Large-scale deep learning framework on FPGA for fingerprint-based indoor localization. *IEEE Access*, 8: 65609–65617.

Liu, K.; Li, Y.; Liu, S.; Tan, C.; and Shao, Z. 2022. Reducing the Label Bias for Timestamp Supervised Temporal Action Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6503–6513.

Luo, D.; Wang, Y.; Yue, A.; and Xu, H. 2022. Weakly-Supervised Temporal Action Alignment Driven by Unbalanced Spectral Fused Gromov-Wasserstein Distance. In *Proceedings of the 30th ACM International Conference on Multimedia*, 728–739.

Luo, J.; Zhang, C.; and Wang, C. 2020. Indoor multi-floor 3D target tracking based on the multi-sensor fusion. *IEEE Access*, 8: 36836–36846.

Menon, V.; Muthukrishnan, G.; and Kalyani, S. 2020. Subspace clustering without knowing the number of clusters: A parameter free approach. *IEEE Transactions on Signal Processing*, 68: 5047–5062.

Ng, Y. B.; and Fernando, B. 2020. Forecasting future action sequences with attention: a new approach to weakly supervised action forecasting. *IEEE Transactions on Image Processing*, 29: 8880–8891.

Qi, S.; Lin, W.; Hong, Z.; Chen, H.; and Zhang, W. 2021. Perceptive autonomous stair climbing for quadrupedal robots. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2313–2320.

Qi, S.; Wu, X.; Wang, J.; and Zhang, J. 2019. Recognition of composite motions based on sEMG via deep learning. In *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 31–36.

Richard, A.; Kuehne, H.; Iqbal, A.; and Gall, J. 2018. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7386–7395.

Sener, F.; and Yao, A. 2020. Unsupervised learning and segmentation of complex activities from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8368–8376.

Sharma, V.; Gupta, M.; Pandey, A. K.; Mishra, D.; and Kumar, A. 2023. A review of deep learning-based human activity recognition on benchmark video datasets. *Applied Artificial Intelligence*, 36(1): 2093705.

Sheng, L.; and Li, C. 2023. Weakly supervised coarse-to-fine learning for human action segmentation in HCI videos. *Multimedia Tools and Applications*, 82(9): 12977–12993.

Sheng, L.; Li, C.; and Tian, Y. 2021. Coarse-to-Fine Loss Based On Viterbi Algorithm for Weakly Supervised Action Segmentation. In *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, 1–6.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Souri, Y.; Fayyaz, M.; Minciullo, L.; Francesca, G.; and Gall, J. 2021. Fast weakly supervised action segmentation using mutual consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6196–6208.

Stein, S.; and McKenna, S. J. 2013. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 729–738.

Sun, X.; Chen, W.; Xiong, X.; Chen, W.; and Jin, Y. 2022a. A Variable Configuration Force Sensor with Adjustable Resolution for Robotic Applications. *IEEE Trans. Ind. Electron.*, 70(2): 2066–2075.

Sun, X.; Wang, C.; Chen, W.; Chen, W.; Yang, G.; and Jin, Y. 2023. A single-actuator four-finger adaptive gripper for robotic assembly. *IEEE Trans. Ind. Electron.*

Sun, X.; Wang, C.; Chen, W.; Yang, S.; He, C.; and Zhi, Y. 2022b. Design and Analysis of a Novel Underactuated Adaptive Gripper for Robotic Assembly. In *IEEE 17th Conference on Industrial Electronics and Applications*, 207–212.

Sun, X.; Yang, Y.; Chen, W.; Chen, W.; and Zhi, Y. 2022c. Grasping Operation of Irregular-Shaped Objects Based on a Monocular Camera. In *International Joint Conference on Energy, Electrical and Power Engineering*, 423–429.

van Amsterdam, B.; Kadkhodamohammadi, A.; Luengo, I.; and Stoyanov, D. 2023. ASPnet: Action Segmentation With Shared-Private Representation of Multiple Data Sources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2384–2393.

VidalMata, R. G.; Scheirer, W. J.; Kukleva, A.; Cox, D.; and Kuehne, H. 2021. 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*, 1238–1247.

Vishwakarma, S.; and Agrawal, A. 2013. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29: 983–1009.

Wang, C.; Luo, J.; Liu, X.; and He, X. 2021. Secure and reliable indoor localization based on multitask collaborative learning for large-scale buildings. *IEEE Internet of Things Journal*, 9(22): 22291–22303.

Wang, C.; Luo, J.; Zhang, C.; and Liu, X. 2020a. A Dynamic Escape Route Planning Method for Indoor Multi-floor Buildings Based on Real-time Fire Situation Awareness. In *2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)*, 222–229.

Wang, H.; and Schmid, C. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, 3551–3558.

Wang, X.; Liu, J.; Mei, T.; and Luo, J. 2023. CoSeg: Cognitively Inspired Unsupervised Generic Event Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, Z.; Gao, Z.; Wang, L.; Li, Z.; and Wu, G. 2020b. Boundary-aware cascade networks for temporal action segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 34–51.