Answers to Reviewer's Comments on "Bringing Online Egocentric Action Recognition into the wild"

Gabriele Goletto Mirco Planamente Barbara Caputo
Giuseppe Averta

March 10, 2023

Dear Editor,

We wish to thank you, and all the Reviewers, for the effort devoted in revising our manuscript, and for the valuable comments provided. The reviews have been precious in assisting us improving the quality of our paper. We made our best to consider all of them in the preparation of a new version of our manuscript (Submission ID: 22-3075, Title: "Bringing Online Egocentric Action Recognition into the wild").

In the following of this document, we provide a detailed point-by-point response to the comments of the Editor and the Reviewers. For Reader's convenience, for each comment we report the Reviewers' text in a grey box, followed by our reply. When appropriate, we also included excerpts of the paper in a white box to show how changes have been implemented.

In the new version of the manuscript, all the changes are highlighted in blue.

Once again, we greatly appreciated all the proactive comments from the Editor and Reviewers. We worked hard to produce a new version of the paper which, we hope, may be positively considered by the Reviewers.

Yours sincerely,

Gabriele Goletto on behalf of all the Authors

Editor

Abstract: The paper was wellreviewed by three reviewers, who agreed that this is an important problem and that the proposed approach had some interesting contributions in tackling the problem. However, several concerns were raised. Chiefly, the lack of citations and comparisons to relevant works and limited ablation studies are a problem. The authors are requested to look at the concerns raised by the reviewers carefully and address them.

Response: We thank the Editor for this summary. In response to the point on the lack of citations and comparisons to relevant works, we revised our paper to include a deeper analysis of existing literature, adding the reference to a number of relevant papers, including [1, 2]. In this manuscript, we clarified better the choice of the dataset used and furthermore introduce new experiments on an expanded version of the dataset to provide further validation of our results.

Additionally, we provide a more extensive ablation study to assess the importance of different components of our proposed approach. Showing the importance to find the best Temporal Window (TW) as a good trade-off between accuracy and inference latency. The dependency of the model with respect to the hyperparameter δ and a good solution to estimate it without tuning it for a specific model or setting.

We have also provided an additional explanation about the research conducted to support our conclusion on the importance of this new line of research and how our benchmark with our proposed solution encourages further research in this direction. We believe that these revisions have strengthened our paper and addressed the concerns raised by the reviewers. Despite the constraints on the number of pages, we assure you that all relevant experimental results will be reported in detail on our project page, along with the code used to generate them, to enable other researchers to reproduce and verify our findings. We hope that the revised manuscript will be considered ready for publication.

Reviewer 245067 (Reviewer 1)

Abstract: This paper defines the requirements for the practical use of egocentric action recognition. These requirements include lightweight models, streaming processing, and online prediction of boundaries. The authors introduce a two-fold aggregator to process overlapping actions in a streaming manner. Results show a successful deployment of efficient egocentric action recognition networks on edge devices.

Comment 1: The authors utilize pre-existing architectures, such as MoViNet, to enable stream processing of frames. It is hard to see the contribution considering that MoViNet already comes with an efficient aggregator to process the full action.

Response:

We thank the Reviewer for this comment, which gives us the opportunity to better clarify the contribution of our paper. The main purposes of this work are: i) to investigate the feasibility of deploying efficient and accurate models - robust to real world constraints - for egocentric action recognition on edge devices, and ii) to provide a (potentially model-agnostic) method for its implementation.
Tackling the first point, the paper contributes with an extensive benchmark on the performance of popular action recognition networks when real-world constraints are posed. This benchmark is novel and, we believe, an interesting contribution that may foster the development of a new line of research targeting a trade off between model accuracy (i.e. mainstream research) and their usability in realistic usecases. Our benchmark demonstrates that, albeit most of the existing models showed promising accuracy and may address some of the constraints listed in the paper, none of them solved them all. For this reason, we worked to develop a method to use (potentially any) existing model under all the aforementioned constraints, which represents the core of this contribution. Although, as correctly mentioned by the Reviewer, MoViNet already implements an aggregator mechanism, this is not necessarily a limitation of our work, as we demonstrated that our approach can be applied also to other feature extractors such as I3D (yet with different results). In addition, our investigation also revealed that MoViNet shows severe drawbacks when it comes to the transition from streaming to an online inference scenario. This is due to a reliance on buffer reset and the absence of a viable method for handling concurrent actions, which we solve in this paper with our contribution to action boundary identification through anomaly detection. The fact that MoViNet is more extensively used in our experiments is motivated by the fact that, to the best of our knowledge, this is to date the best option to deploy on edge devices, and motivates research on the development of tiny models for egocentric action recognition.

In other words, we strongly believe that the core, and the point of strength, of this contribution is to provide a way to use (potentially) any features extractor under real world constraints, enabling the update of our approach with new models in the future.
We acknowledge that the first version of the manuscript failed in transferring this concept to the reader, and therefore we edited the Intro to better highlight the contribution of our work, as reported in the box below.

– In Page 2 left column, end of Intro section. –
To summarize, this paper contributes with:

- the definition of a new setting of FPAR in the wild, which encourages researchers to develop applications-aware solutions;
- a benchmark of popular action recognition models for real-world application in FPAR;
- a method to enable the use of existing features extractors to achieve efficient yet accurate action recognition under constraints, exploiting an anomaly detection strategy to localize the boundary of the actions and a two-fold aggregator solution to deal with concurrent actions in a continuous stream;
- an analysis of performance on an edge device, opening interesting perspectives for on-board intelligence.

Comment 2: Detecting boundaries by identifying troughs in the similarity between frames in an online manner (without parameters or training) has been introduced in [35 in paper]. It is not clear whether the authors use anomalies (unknown class) to detect boundaries or distance between features; both are introduced in the same section under DBL. Using [35 in paper] framework on supervised representation (trained in domain with labels) is expected to work much better than the unsupervised representation in [35 in paper]; this is not a significant contribution.

Response:

In this work, we propose a solution for detecting boundaries in online fashion video sequences by leveraging existing action recognition models. We acknowledge the similarity between our approach and ABD [3], as both methods use a comparison between features of different frames to detect boundaries. Our main goal is to develop a pipeline that facilitates the easy repurposing of existing action recognition models for online fashion applications. We believe that the ABD approach serves as a valid starting point for our benchmark.

However, our proposed solution addresses two limitations present in the ABD approach: i) a non-negligible latency (50 frames in the worst case, i.e. ≈ 1.6 seconds) introduced by the NMS window during the aggregation step, and ii) an increased memory budget caused by the need to store two different models (one fine-tuned on the target dataset and one pre-trained on a large scale dataset, i.e. kinetics). These limitations are important consideration for real-world applications and our approach addresses them effectively.

Our solution also offers a more solid motivation from the perspective of unsupervised anomaly detection. Our method is able to detect the beginning or end of known classes through the same mechanism, which is particularly useful in cases where two known actions are interleaved by "no-activity" or an unknown action. Standard models are typically trained on trimmed actions only, where all frames in a clip depict the same action, which results in the model mapping frames of a given action to the same region of the feature space. When an input clip contains two consecutive actions, the transition at the frame level between the two actions produces a large variation in the feature representations [4], which may be regarded as an anomaly [5]. A qualitative representation of this behavior is reported in Figure 1 (other than the video submitted jointly with the paper).

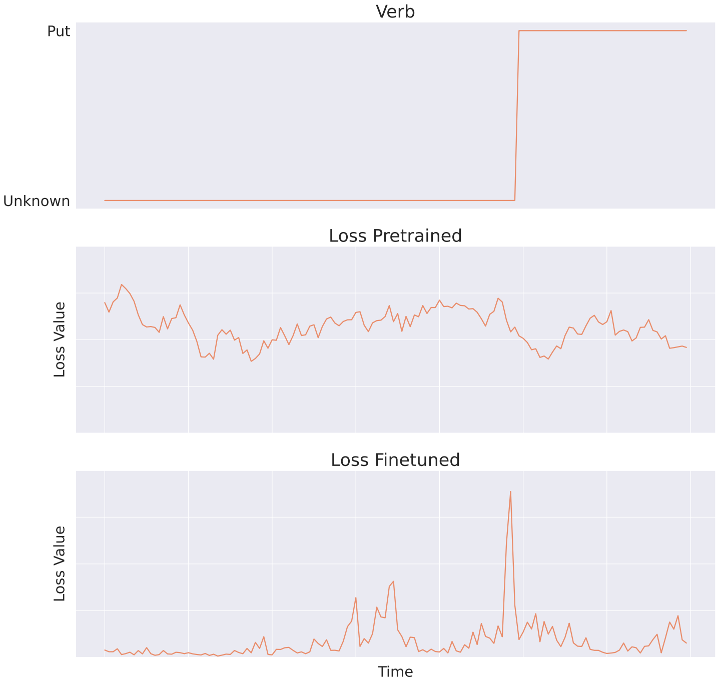

Figure 1: Plot of the Mean Squeer Error (MSE) among the I3D model's features for both the pretrained and finetuned versions.

Lastly, although the similarities between our solution and ABD, our approach demonstrates its superiority in fine-grained egocentric action detection. We hypothesize that the ABD limit in this context is attributed to the fact that the network does not have knowledge of fine-grained actions, resulting in it being highly biased towards the environment change, i.e. when the user changes location in the kitchen, rather than the motion performed during the fine-grained action itself.

Comment 3: The main contribution in the architecture seems to be using two aggregators to account for overlapping actions.

Response: We kindly disagree with the Reviewer on this comment. As already partially discussed in Comment 1, how to handle concurrent actions through a double aggregators is only a part of our contribution, which includes the definition of a new setting for egocentric action recognition, an extensive benchmark of existing models under real use constraints and a method to use existing models to solve the task fitting all the constraints.

Comment 4: Relevant literature [2,3] on self-supervised detection of events/boundaries from untrimmed streaming/long video is not mentioned.

Response: Thank you for pointing out that we missed some relevant literature on self-supervised detection of events/boundaries from untrimmed streaming/long video. We revised our Related Works section to include also the suggested papers.

Comment 5: Additionally, approaches like UnWeaveNet[4] discuss threads of actions that pause and resume throughout the video, showing that actions can be weaved together. Please discuss how the aggregator implementation can be modified to handle frames that are not consecutive and how this approach can account for an unknown number of weaved actions.

Response:

We thank the Reviewer for this comment, which gives us the opportunity to better explain our contribution. There are indeed some points of contact between our work and UnWeaveNet[4], since both deals with the problem of untrimmed videos. However, our method focuses on detecting and recognizing individual actions such as ”take,” ”put,” and ”open,” while UnWeaveNet [4] separates all actions within a video and then i) identifies daily activities like making toast and preparing coffee and ii) when they are started or restored. The concept of restore is not easily transferred to our method of fine-grained action recognition, as it is not taught during the standard training for action recognition which only learns the actions without any information about the activity being performed.

The reviewer's comment suggests potential opportunities for future research to integrate our method with UnWeaveNet [4]. By combining UnWeaveNet's concept with our online fine-grained action recognition, we can learn to distinguish the actions performed in each sub-activity and make necessary adjustments to make UnWeaveNet more efficient and fast, so it can be used on real-time devices.

Comment 6: Why is the streaming approach tied to egocentric video? It seems processing video in a streaming manner is a practical approach for all video/audio processing applications (e.g., surveillance). The practicality of streaming datasets/approaches for videos should be decoupled from the egocentric type of videos. In other words, other papers have discussed the streaming problem, what makes the egocentric video more challenging or different from other video/audio types that it requires its own custom streaming approach.

Response:

We concur with the Reviewer that the approach for analyzing online videos has relevance for third-person input applications, such as surveillance. However, we contend that working with egocentric videos presents a more demanding set of constraints, both in terms of available resources and the challenges imposed by this type of data.

Devices that collect third-person videos are typically static and can, in theory, be connected to a power source or have larger capacity batteries. This is not the case with wearable devices, which are subject to more severe power consumption constraints. Additionally, as stated in [6], the wild movement of the camera and lack of context in egocentric videos make it challenging to recognize actions with the same level of accuracy as third-person vision solutions. The ego-motion and domain changes that result from user movement are challenges that we believe are vital to consider in research such as ours, and are primarily found in egocentric videos.

Furthermore, the importance of privacy is another crucial factor that drives the motivation of our research in egocentric videos. In contrast to third-person perspectives, where privacy is typically protected by using blurring techniques, the utilization of first-person video creates a considerable privacy concern [7]. This is because the recorded data includes personal characteristics such as the user's gait or other biometric information, which cannot be obscured by traditional blurring methods. By investigating the feasibility of implementing action recognition models and other tasks on the device, we can ensure the preservation of user data on the device, thereby augmenting the level of security and privacy as opposed to transmitting the data to external servers.

In conclusion, although the analysis done is easily extendable to third-person videos, we focused on the egocentric scenario because it is more challenging and, consequently, the one we believe leads to the development of more robust solutions.

Comment 7: The terminology here is important. Please clarify in the text that all tasks (offline, streaming, and online) refer to inference only. So these should be renamed as offline inference, streaming inference, and online inference to avoid confusion with streaming approaches that do training and inference together.

Response: Thank you for your feedback on the terminology used in our paper. We apologize for any confusion caused by our use of the terms “offline”, “streaming”, and “online” to describe the tasks in our work.

We have now revised the manuscript to reflect this, and have renamed the tasks as “offline inference”, “streaming inference”, and “online inference” to avoid any confusion with streaming approaches that involve both training and inference.

Thank you for bringing this to our attention, and we hope that these changes will improve the clarity of our manuscript.

Comment 8: In Section III, the authors mention, “the goal is to find a good trade-off between: i) the amount of information needed as input to properly encode the temporal information”; however, important ablations on the amount of input information needed are not provided. It is a good idea to do ablations on the frame window size (TW) as this seems to be a significant factor possibly affecting the performance.

Response: We thank the reviewer for pointing out the absence of ablation on the size of the temporal window given as input to the network itself. To address this problem, we tested a wide range of values for the hyperparameter TW, training and consequently testing the networks with these. We report the accuracy obtained in Table 1 for both I3D and MoViNet-A0 (without streaming buffer).

From the results, we observe that in both the Seen and Unseen scenarios, although using TW= 16 is not always the optimal choice in terms of accuracy for the Offline inference setting, this represents the best trade-off between computational cost and performance. In these regards, we report the effect produced by an increase of the temporal window on MACs. Interestingly, since MACs are linearly correlated with TW, increasing the window only to improve performance is not a viable solution for deployment on edge devices. In addition, large temporal windows do not help in the Streaming inference scenario where a big TW causes a bigger overlap among clips as demonstrated by the results

Table 1: EPIC-Kitchens performance. Top-1 *mean* accuracy (%), ablation on Temporal Window size.

EPIC-Kitchens					
Network	Mode	TW	Seen	Unseen	MACs
I3D	*Offline*	8	60.98	39.17	$130e^8$
I3D	*Offline*	16	67.08	42.42	$270e^8$
I3D	*Offline*	24	67.81	44.92	$410e^8$
I3D	*Offline*	32	66.66	45.11	$550e^8$
I3D	*Offline*	48	65.06	43.21	$830e^8$
MoViNet-A0	*Offline*	8	59.87	36.34	$3.9e^8$
MoViNet-A0	*Offline*	16	64.17	40.68	$7.7e^8$
MoViNet-A0	*Offline*	24	64.89	42.36	$11e^8$
MoViNet-A0	*Offline*	32	65.31	43.11	$15e^8$
MoViNet-A0	*Offline*	48	63.36	42.39	$23e^8$
I3D	*Streaming*	8	61.81	38.59	$130e^8$
I3D	*Streaming*	16	63.38	40.57	$270e^8$
I3D	*Streaming*	24	60.06	40.83	$410e^8$
I3D	*Streaming*	32	56.4	40.78	$550e^8$
I3D	*Streaming*	48	52.07	39.05	$830e^8$
MoViNet-A0	*Streaming*	1	62.24	39.59	$0.47e^8$

where 16 seems to be the best clip temporal length. Finally, using a temporal window of 16 frames makes the results directly comparable with various works in action recognition using I3D model [8, 9, 10], making the benchmark itself more relevant for the research community.

With regard to MoViNet-A0 with a streaming buffer, however, it would not make sense to evaluate a variable TW as this would greatly worsen the performance of the model. Considering the streaming scenario where the buffer continues to accumulate information on past features, using a TW greater than one causes both the input of redundant information into the buffer (resulting in performance degradation) and an increase in inference latency.

Comment 9: It is unclear why the streaming results for MoViNet reports (52% and 35% for seen and unseen respectively) in Fig. 7, but reports (62.24% and 39.59% for seen and unseen respectively) in table II. The streaming case assumes knowledge of the boundaries, so the results should be the same, correct?

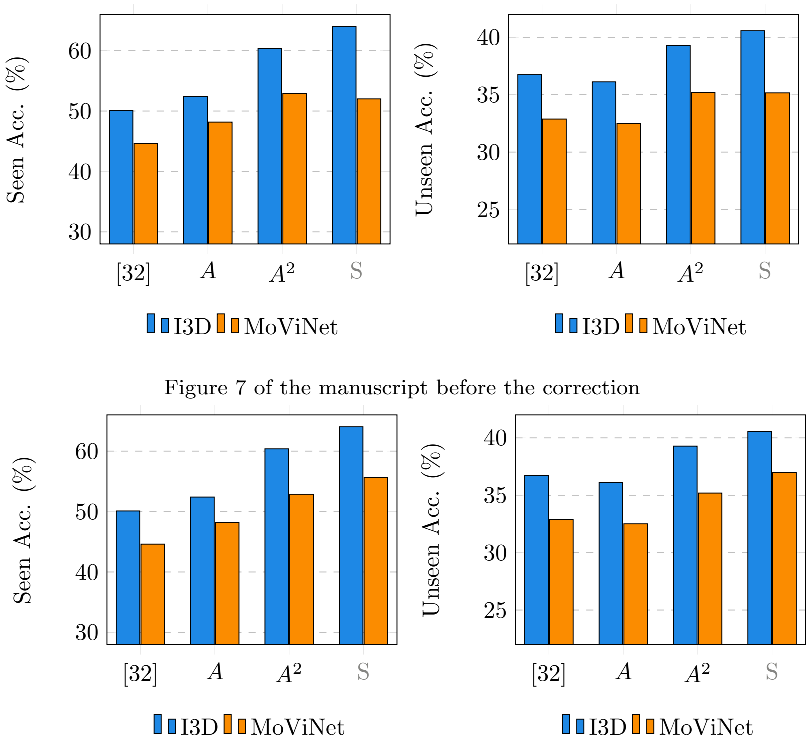

Figure 7 of the manuscript before the correction

Figure 7 of the manuscript after the correction

Figure 2

Response:

We thank the Reviewer for raising this issue, which indeed is a typo in the first version of the manuscript. We take this opportunity to mention that we corrected the following values (without affecting the findings of the paper):

- The results for MoViNet without streaming buffer in Table I
- The number of MACs relating to I3D in Table III

However, we would like to point out that, although the untrimmed streaming (i.e., supervised on boundaries) model accuracy for MoViNet is indeed higher than the first version (see Figure 2 for the differences), the two results mentioned by the Reviewer do not necessarily correspond. Indeed, although in both cases the boundaries of the actions are known a-priori by the model, in the first case (Tab.2) the validation is performed in trimmed mode, while in the latter (Fig. 7) it is performed in untrimmed mode. These represent a different type of supervision because in the trimmed setting the buffer is always reset exactly at the beginning of each action (since trimmed actions are temporally juxtaposed).

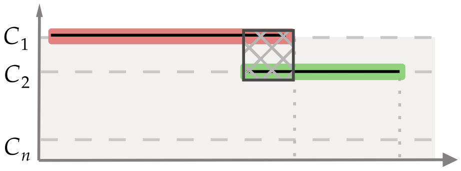

Figure 3: A schematisation of two overlapping actions

In the untrimmed setting, on the other hand, two actions may overlap each other and the buffer is still reset at the beginning/end of each action. This may result in cases where the buffer is partially filled with information about action C_1 in the example in Figure 3 (which starts before action C_2). However, as soon as action C_2 begins, overlapping with C_1, the buffer is emptied and loses all information relating to C_1, thus incorrectly predicting it.

In the new version of the manuscript we report the correct values for both experiments. Interestingly, for the untrimmed case, we still get lower performance w.r.t. trimmed videos, with 55.6% and 37% at supervised and cross domain, respectively, while in the trimmed case we obtain 62.24% and 39.59% on seen and unseen kitchens, correspondingly.

Such gap between performances in trimmed and untrimmed settings suggests that overlapping actions are not a rare event, and this condition should be considered. In our methods, we introduced a two-fold aggregator mechanism to minimize the loss of information caused by model buffer reset in case of overlapping actions. In the streaming case (i.e. when the knowledge of action boundaries is available), the two buffers store knowledge of two different actions separately, without loss of information. Our results demonstrate that this is a viable solution to close the accuracy gap between supervised trimmed and untrimmed scenarios, achieving 60.97% and 39.42% (seen and unseen respectively) versus 62.24% and 39.59% (seen and unseen in the trimmed scenario).

Comment 10: When reporting results for [35 in paper] in Fig. 7, was NMS used to remove duplicated boundary detections? [35 in paper] is re-implemented on epic-kitchen, so it is important to list the hyperparameter values used and any modifications to the original paper implementation details.

Response: Thank you for pointing out this omission. In our experiments, we used the complete pipeline mentioned in Section 3.4 of [3], and therefore we included the NMS to remove duplicated boundary detections. In particular, we opted for design choices as close as possible to the original implementation but keeping in mind the efficiency and low latency required in the setting proposed. Indeed, we adopted average filtering instead of gaussian one as they showed consistent performance in the experiments of [3]. The size of the filter and the one of the NMS window are equal and set to 50 (i.e., $L = k + 1 = 50$). We chose it by performing the same ablation as shown in Figure 3 of [3] and selecting the best performing L value while keeping the actual trade-off between accuracy and latency in mind. Of note, this is a critical hyperparameter since the largest is the NMS window, and the highest is the worst-case latency for boundary detection. We clarified this detail in the paper by adding in section VI (page 7):

> – In Page 7 left column, end of Section VI –
> For ABD, we used the original online implementation, with both NMS and filter windows size equal to 50.

Comment 11: It says in the text that "rapid changes in background, environment, perspective, and illumination" from using a head-level worn sensor is typically referred to as domain shift. This may not be accurate; a domain shift is usually training in one environment or for one task and testing in another. If the model was trained using the same headlevel worn sensor and in the same environment, is it considered domain shift?

Response:

Thank you for bringing this to our attention. We apologize for any confusion caused by our original submission and have made the necessary changes to ensure that the text is clear and easy to understand.

To provide further clarity, with domain shift we refer in general to any change in the distribution of input data or in the conditions under which a model is applied, which can lead to decreased performance. This can occur when the model is tested on data different from the data it was trained on.

In the context of using a head-level worn sensor, changes in background, environment, perspective, and illumination could be considered a form of domain shift if they occur between training and testing. However, if the model was trained using the same head-level worn sensor and in the same environment as it will be tested, it is not considered a domain shift.

It is crucial to point out, however, that the assumption that the sensor and environment will remain unchanged during training and testing is unrealistic. This is especially true in the egocentric scenario, where there is a high degree of variability in the data due to the user's movement, perspective, lighting, location, and the presence of other people and objects in the environment. As a result, it is very difficult to have a training set that generalizes to every aspect of real life in this scenario.

– In Page 2 right column, end of first part of Section III –

It is also worth reporting that, because the sensor is worn by the user - usually at the head level - it records data with a high degree of variation produced by rapid changes in environment, perspective, and illumination as in Fig [2]. Input variability can cause a difference in the distribution of data between the training and testing phases. This results in a problem known as *environmental bias* or *domain shift* that can negatively impact the performance of the model.

Reviewer 245069 (Reviewer 2)

Abstract: This paper discusses the challenges of egocentric videos deployment for realistic applications and solutions to tackle some problems. The authors discuss about some constraints such as model portability, computational power for real-time inference on edge device, domain shift etc. and proposes some solutions which can work on top of existing architecture. This paper presents detailed experiments and analysis for the proposed solutions and challenges using the Epic-Kitchen dataset. The paper also presented a benchmark for realistic deployment of egocentric videos and shed light on the feasibility of deploying a model on a smaller device which can work for real-time egocentric videos with very low energy. [...] Some places in the paper are difficult to follow and seems a little ambiguous such as:

Comment 12: we analyze the feasibility of deploying an egocentric vision model on a budget, opening interesting perspectives for on-board intelligence. [...]

Response: In our analysis, we meticulously evaluated the potential costs and benefits of implementing an egocentric vision model, which is a type of machine learning model capable of processing visual information recorded from the perspective of an individual. After conducting a thorough analysis, we have determined that deploying this model within a budget-conscious framework is feasible, meaning that we are mindful of the available resources and seek to use them efficiently. The results of this research could have significant ramifications for the advancement of on-board intelligence. Specifically, the utilization of an egocentric vision model has the potential to augment the capabilities of a variety of systems designed for use while in motion.

– In Page 2 left column, end of Intro section. –
an analysis of performance on an edge device, opening interesting perspectives for on-board intelligence.

Comment 13: wild lies on how data input are structured. [...]

Response: We apologize for any confusion caused by the original sentence in the submission. To clarify, the sentence was referring to the fact that the action recognition protocol is designed to work with data that is structured in a specific way, specifically video data that has been "trimmed" to only include the portion of the video where a specific action is present. This means that the data input to the model has been supervised in some way to identify the beginning and end of the action.

However, in a real-world deployment of this technology, it is important to consider the limitations of using a standard action recognition approach in situations where the data input may not be structured in this way, i.e., when working with continuous streams of data where there is no clear information about the start or end of an action. We have made the necessary changes to the text to ensure that it is clear and easy to understand.

– In Page 3 left column, end of first part of Section III –
...lies in the intrinsic untrimmed nature of input data.

Comment 14: requiring the samples length information as for uniform sampling to obtain a video level prediction.[...]

Response: To clarify this sentence we shortly describe the differences between uniform vs dense and clip vs video level sampling as depicted in [11]. Specifically, at inference time to calculate the output for the current action, the values obtained from several different clips are averaged. Although the sampling used to obtain these clips is dense (e.g., 16 consecutive frames), at video level, the clips themselves are uniformly distributed within the action. This implies knowing the total duration of the video in order to uniformly sample the clips within.

– In Page 3 right column, Section III A –
Indeed, the final prediction is usually obtained by averaging the predictions of different equidistant clips over the whole video, performing video level uniform sampling, i.e. requiring the sample's length information.

Comment 15: To relax also the assumption on action boundaries etc.[...]

Response: To further clarify this sentence, let us explain in more detail the difference between streaming and online inference. Although both settings are based on continuous frame processing, when we speak of streaming inference we assume that the model still knows the temporal limits (i.e., beginning and end) of the analyzed actions whereas in online inference even this type of supervision is lacking.

– In Page 3 right column, end of Section III A –
Removing the supervision on action boundaries as well...

Comment 16: The inference time analysis of different models and variants might add some insight for real-world deployment of egocentric videos

Response: We agree with the Reviewer that inference time is crucial for our application, which is the reason why we reported it in Tab. III (column ”Latency”). To better clarify that, with latency, we refer to inference time as we explicitly mentioned in the caption of the table.

Reviewer 245073 (Reviewer 4)

Abstract: This paper evaluates the performance of state-of-the-art egocentric action recognition methods in real-world applications. In particular, it considers different aspects such as hardware restrictions, cross-domain scenarios, and online inference on untrimmed data. Moreover, it proposes a two-fold aggregator to "switch" between the actions or "detect" the action boundaries. Experiments are conducted on a subset of the EPIC-Kitchen-55 dataset (i.e., videos from 3 kitchens, out of the 32 kitchens available in the dataset).

Comment 17: The experiments are conducted only on a small dataset (i.e., 3 out of 32 kitchens in EPIC-Kitchen-55)

Response:

We understand your concern regarding the limited number of kitchens used in our experiments, however, we politely disagree with your assessment that this is a small dataset. The three kitchens were specifically selected because they contain the most labeled samples. Furthermore, it may be relevant to note that the setting used in this paper is the standard de-facto for cross-domain analysis for first person vision, as demonstrated by the references provided in our manuscript [8, 12, 13] and also [10, 14, 15, 16, 17, 18, 19]. In addition, the size of our dataset is comparable to other datasets used for egocentric action recognition or cross-domain settings, as summarized in Table 2 and we are confident that this setting is representative of the whole dataset as (i) it keeps the division into multiple domains and (ii) the actions maintain the strong inter-class unbalance of the original dataset, as shown in [8] (Fig. 4).

Table 2: Different popular dataset.

Datasets			
Dataset	Year	Modalities	Samples
GTEA[20]	2011	RGB	525
FPAH[21]	2018	RGB-Depth	1,175
EGTEA Gaze+[22]	2018	RGB+Gaze	10,325
EPIC-Kitchens 3 kitchens[23]	2018	RGB+Audio+Event	10,094
EPIC-Kitchens 7 kitchens[23]	2018	RGB+Audio	12.427

Furthermore, we chose this dataset because it contains rich multi-modal information that is interesting to study in the online action recognition setting, including audio and event data. For example, the extension introduced in [13] offers a valuable alternative to standard RGB data, with the added benefits of lower power consumption and suitability for wearable devices.

However, we understand that the Reviewer may have some concerns on the validity of our results and, for this reason, we performed additional analysis by adding four more kitchens to our dataset, bringing the total number of samples to 12427. The results of testing our model with these new kitchens are reported in Table 3, and they demonstrate that the overall conclusions made in our paper remain consistent even in this expanded scenario. The final performance also remains similar, further supporting the validity of our benchmark.

Table 3: EPIC-Kitchens new kitchens performance. Top-1 *mean* accuracy (%), models trained all D_1, D_2, D_3 separately and tested on the new kitchens.

	EPIC-Kitchens new kitchens		
Network	Mode	Sampling	Unseen
I3D	*Offline*	D 16x5	41.58
MoViNet-A0	*Offline*	D 16x5	38.53
I3D	*Streaming*	All Stream	38.74
MoViNet-A0	*Streaming*	All Stream	36.12
I3D	*Online untrimmed - DBL+A*	All Stream	31.33
I3D	*Online untrimmed - DBL+A*^{2}	All Stream	36.44
MoViNet-A0	*Online untrimmed - DBL+A*	All Stream	30.98
MoViNet-A0	*Online untrimmed - DBL+A*^{2}	All Stream	32.91

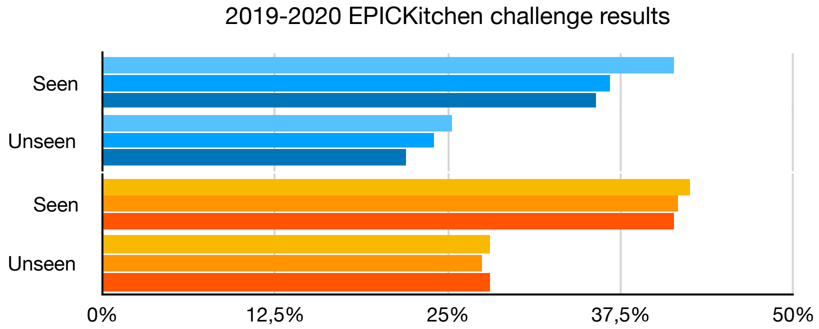

Figure 4: A comprehensive analysis of the top 3 winning methods of the EPIC-Kitchen challenges in 2019 [24] and 2020 [25] competitions. The evaluations were performed on both *Seen* and *Unseen* kitchens.

It is worth noticing that our analysis focuses on a setting called single source, in which training is conducted on a single kitchen and testing is conducted on a different one. However, we would like to clarify that the domain shift problem we address in our paper, in which the training and testing datasets come from different domains, is not limited to this single source setting. Indeed, the popular epic-kitchen challenge has demonstrated that even when training on a large dataset comprising multiple kitchens and testing on a different group, ensemble solutions with a high number of parameters can still struggle with cross-domain issues, as shown in Figure 4 adapted from [12], where the top three competitors of EPIC-Kitchen challenges still suffered of $\approx 10\%$ of performance drop in the unseen scenario. From this standpoint, we believe that it would be interesting to investigate the use of online Unsupervised Domain Adaption (UDA) or Test Time Adaptation (TTA) approach in future work as a potential solution to this problem.

We hope that the new analysis included in this rebuttal may help in clarifying our contribution and the relevance of the dataset used for the purposes of our study.

– In Page 5 left column, Section V. Implementation –
In our experiments, we utilize the top three kitchens with the most labeled samples from the EPIC-Kitchens-55 dataset [23]. These kitchens are referred to as D1, D2, and D3. We have chosen this specific setting as it is the standard and widely used dataset for cross-domain analysis in first-person perspective [8], and it also provides rich multi-modal information, including audio and event data [13], which can be beneficial for further analysis. Additionally, the difficulties in this dataset arise not only from the significant domain shift among different kitchens, but also from imbalanced class distribution both intra- and inter-domain.

Comment 18: The contributions are limited (i.e., the two-fold aggregator and the evaluations). In particular, the two-fold aggregator is merely a weighted sum of two aggregators (i.e., Eq.1) and is likely not able to handle three or more overlapping actions.

Response:

We kindly disagree with the Reviewer on this comment. As already partially discussed in Comment 3 from the first Reviewer, how to handle concurrent actions through double aggregators is only a part of our contribution.

Indeed, the main purposes of this work are: i) to investigate the feasibility to deploy efficient and accurate models - robust to real-world constraints - for egocentric action recognition on edge devices, and ii) to provide a (potentially model-agnostic) method for its implementation.

Tackling the first point, the paper contributes with an extensive benchmark on the performance of popular action recognition networks when real-world constraints are posed.

This benchmark is novel and, we believe, an interesting contribution which may foster the development of a new line of research targeting a trade off between model accuracy (i.e. mainstream research) and their usability in realistic use-cases. Our benchmark demonstrates that, albeit most of the existing models showed promising accuracy and may address some of the constraints listed in the paper, none of them solved them all. For this reason, we worked to develop a method to use (potentially any) existing model under all the aforementioned constraints, which represents the core of this contribution.
We strongly believe that the core, and the point of strength, of this contribution is to provide a way to use (potentially) any features extractor under real world constraints, enabling the update of our approach with new models in the future. We acknowledge that the first version of the manuscript failed in transferring this concept to the reader, and therefore we edited the Intro to better highlight the contribution of our work, as reported in the box below.

The problem of overlapping action is particularly relevant in the field of action recognition in videos, where the goal is to correctly identify and label the actions that are occurring in a given video sequence. This can be a challenging task, especially when the actions are fine-grained and occur in a continuous stream. The difficulty arises from the fact that it is often difficult to detect and establish the boundaries of these fine-grained actions, even from a human perspective.

For example, in a scenario where a user takes an object to perform some task, it can be hard to define the specific point in time that separates the action of "take" the object and the beginning of the second action. This difficulty in determining the start and end of fine-grained actions in a continuous stream is what we refer to as the problem of overlapping action.

Considering this above issue, we highlight that attempting to eliminate overlapping by using the end of one action as the start of the consecutive one, negatively impacts overall performance. This is because it reduces the duration of at least one of the actions, making it harder to recognize. To address this issue, our solution, the double buffer, allows the model to begin identifying the next action without interrupting the previous one, thus improving recognition and overall performance.

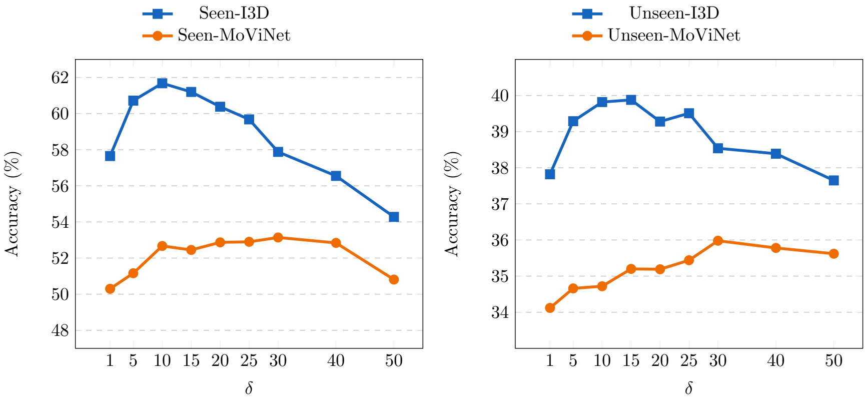

Figure 5: The effect of the delay parameter δ on the three kitchen settings was investigated. We report Top-1 *mean* accuracy (%), over all $D_i \rightarrow D_j$ combinations on both seen and unseen test sets in *online-untrimmed* setting.

– In Page 2 left column, end of Intro section. –

To summarize, this paper contributes with:

- the definition of a new setting of FPAR in the wild, which encourages researchers to develop applications-aware solutions;
- a benchmark of popular action recognition models for real-world application in FPAR;
- a method to enable the feasibility to use existing features extractors to achieve efficient yet accurate action recognition under constraints, exploiting an anomaly detection strategy to localize the boundary of the actions and a two-fold aggregator solution to deal with concurrent actions in a continuous stream;
- an analysis of performance on an edge device, opening interesting perspectives for on-board intelligence.

Comment 19: The effect of the delay hyperparameter δ on the accuracy is not studied.

Response:

We agree with the Reviewer that we did not provide sufficient details on how the hyperparameter regulating the delay δ was chosen. In our implementation, we estimated δ directly from the dataset by calculating the average number of frames (at 30 Hz) that presented an overlap of at least two actions. This was performed on a subset of kitchens different than the ones we used in our experiments, to avoid introducing a bias that may favour our results, and set to 20. To ablate the value of δ following the Reviewer's suggestions, we performed a set of experiments varying the value in the range $[1, 50]$. Results, depicted in Fig. 5, suggest that the range $[10, 30]$ presents comparable results, while performance drops outside this range. We clarified our tuning of δ also in the paper by adding:

– From Page 5, left column (Implementation Details): –
For the two-fold aggregator implementation, the value of δ was estimated directly from the dataset (a subset of kitchens from [20] not used in this paper) by calculating the average number of frames (at 30 Hz) that presented an overlap of at least two actions.

Comment 20: Furthermore, as mentioned above, the evaluations are limited to a small dataset. It would be more beneficial and convincing to the readers if more and larger datasets are used and more discussions and suggestions regarding handling hardware restrictions, cross-domain scenarios, and online inference on untrimmed data are provided.

Response: As extensively discussed in Comment 17, the setting used in this paper is the standar de-facto for research on cross-domain egocentric action recognition, and its size is comparable with other datasets used for egocentric action recognition or cross-domain settings, as summarized in Tab. 2. However, to provide further evidences on the validity of our results, we expanded the pool of kitchens considered, resulting in an higher number of cross-domain shifts. Please refer to Comment 17 for details on these results.

With regards to the hardware restrictions, we concur with the reviewer that the problem of handling these limitations, cross-domain scenarios, and online inference on untrimmed data is of great relevance to this manuscript. We have added a discussion of this topic in the Conclusions section of the paper, albeit constrained by the page limit of the Letters. Furthermore, more experiments are presented in Comment 8, which indicate the impact of TW size on performance and MACs values, as these are some of the factors to consider during device deployment.

– In Page 7, end of Conclusion section: –
We believe that anomaly detection-based strategies and aggregator solutions represent powerful tools to enable video processing on the edge, and that this task may significantly benefit from Unsupervised Domain Adaption (UDA) or Test Time Adaptation (TTA) techniques to properly address the challenges that arises when tackling scene understanding from untrimmed videos. Future works will consider such challenges, with the development of method for the continuous adaptation of the model during the untrimmed video processing.

References

[1] R. Mounir, R. Gula, J. Theuerkauf, and S. Sarkar, "Spatio-temporal event segmentation for wildlife extended videos," in *International Conference on Computer Vision and Image Processing*, pp. 48–59, Springer, 2022.

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

[3] Z. Du, X. Wang, G. Zhou, and Q. Wang, "Fast and unsupervised action boundary detection for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[4] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041–1049, 2017.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, 2009.

[6] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Egocentric vision-based action recognition: A survey," *Neurocomputing*, vol. 472, pp. 175–197, 2022.

[7] D. Thapar, A. Nigam, and C. Arora, "Anonymizing egocentric videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2320–2329, 2021.

[8] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 122–132, 2020.

[9] J. Choi, G. Sharma, S. Schulter, and J.-B. Huang, "Shuffle and attend: Video domain adaptation," in *European Conference on Computer Vision*, pp. 678–695, Springer, 2020.

[10] D. Kim, Y.-H. Tsai, B. Zhuang, X. Yu, S. Sclaroff, K. Saenko, and M. Chandraker, "Learning cross-modal contrastive features for video domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13618–13627, 2021.

[11] C.-F. R. Chen, R. Panda, K. Ramakrishnan, R. Feris, J. Cohn, A. Oliva, and Q. Fan, "Deep analysis of cnn-based spatio-temporal representations for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6165–6175, 2021.

[12] M. Planamente, C. Plizzari, E. Alberti, and B. Caputo, "Domain generalization through audio-visual relative norm alignment in first person action recognition," in *WACV*, January 2022.

[13] C. Plizzari, M. Planamente, G. Goletto, M. Cannici, E. Gusso, M. Matteucci, and B. Caputo, "E2 (go) motion: Motion augmented event stream for egocentric action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19935–19947, 2022.

[14] J. Lv, K. Liu, and S. He, "Differentiated learning for multi-modal domain adaptation," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1322–1330, 2021.

[15] M. Plananamente, C. Plizzari, and B. Caputo, "Test-time adaptation for egocentric action recognition," in *International Conference on Image Analysis and Processing*, pp. 206–218, Springer, 2022.

[16] L. Yang, Y. Huang, Y. Sugano, and Y. Sato, "Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14722–14732, 2022.

[17] P. Wei, L. Kong, X. Qu, X. Yin, Z. Xu, J. Jiang, and Z. Ma, "Unsupervised video domain adaptation: A disentanglement perspective," *arXiv preprint arXiv:2208.07365*, 2022.

[18] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, and H. Chai, "Spatio-temporal contrastive domain adaptation for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[19] A. Sahoo, R. Shah, R. Panda, K. Saenko, and A. Das, "Contrast and mix: Temporal contrastive video domain adaptation with background mixing," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23386–23400, 2021.

[20] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *CVPR 2011*, IEEE, June 2011.

[21] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[23] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 720–736, 2018.

[24] D. Damen, W. Price, E. Kazakos, A. Furnari, and G. M. Farinella, "Epic-kitchens - 2019 challenges report." https://epic-kitchens.github.io/Reports/EPIC-Kitchens-Challenges-2019-Report.pdf, 2019.

[25] D. Damen, E. Kazakos, W. Price, J. Ma, and H. Doughty, "Epic-kitchens-55 - 2020 challenges report." https://epic-kitchens.github.io/Reports/EPIC-KITCHENS-Challenges-2020-Report.pdf, 2020.