Diagnosing Error in Temporal Action Detectors (Supplementary Material)

Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem

King Abdullah University of Science and Technology (KAUST), Saudi Arabia
http://www.humamalwassel.com/publication/detad/
{humam.alwassel, fabian.caba, victor.escorcia, bernard.ghanem}@kaust.edu.sa

This document provides supplementary information to the main findings and methodology described in the paper. The document is organized as follows: **(i)** Section 1 analyzes the impact of using multiple annotations for the evaluation of action localization algorithms (mentioned in Section 8); **(ii)** Section 2 covers additional details and analysis in ActivityNet [3]; and **(ii)** Section 3 discloses the results of our analysis in THUMOS14[4].

1 New Evaluation Metric

Considering that multiple annotators disagree on the starting and ending time of some instances, it is pertinent to examine an alternative evaluation that takes this observation into account. Here, we consider a flexible evaluation which assumes that all annotators are equally right. In that sense, we consider that a prediction matches a ground truth instance, described in terms of a set of annotations $\mathcal{A}$, if its tIoU with one element of the set $\mathcal{A}$ exceeds a given threshold.

$$\text{TP}_{p,\mathcal{A}_k} = 1_{\max\limits_{a \in \mathcal{A}_k} \text{tIoU}(p,a) \geq \alpha} \tag{1}$$

where p represents a given prediction, $\mathcal{A}_k$ the set of annotations for the k-th instance, α the overlap threshold of interest, and 1 represents the indicator function. $\text{TP}_{p,\mathcal{A}_k}$ equal to 1 represents that the prediction p is a true positive for

Table 1: Average-mAP$_N$ performance on ActivityNet as we vary the number of annotations (k) per ground truth instance. The average-mAP$_N$ improves across all methods and the gap in performance between methods increases as k increases.

	Average-mAP$_N$ (%) Using k Annotations per Instance			
Method	1	2	3	4
SC	33.92	46.04	52.92	57.51
CES	32.24	43.31	49.64	53.86
IC	32.14	41.80	47.18	50.52
BU	17.26	25.90	31.79	36.22

Error Type	Overall Error Breakdown (%) $k=1$	$k=4$	Diff
TP	5.68	6.98	+1.30
DD	12.83	21.99	+9.16
WL	0.70	1.03	+0.33
LOC	63.19	59.04	−4.15
CON	2.21	2.04	−0.17
BG	15.39	8.91	−6.48

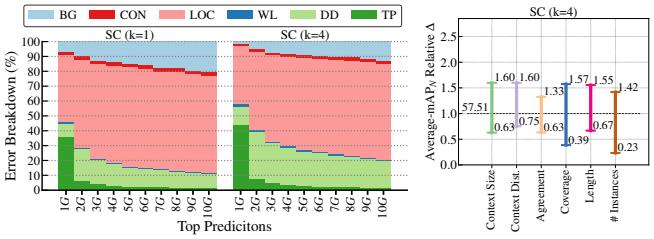

Fig. 1: BG=Background Error; CON=Confusion Error; LOC=Localization Error; WL=Wrong Label Error; DD=Double Detection Error; TP=True Positive.

the k-th ground truth instance. Similarly to the evaluation protocol described in the Section 2 of the main paper, once a ground truth instance is matched to a given prediction any prediction matching that instance is considered a double detection (DD). In this case, a double detection is a prediction that match any element of the set $\mathcal{A}$ associated with an instance already matched.

Table 1 shows the results for all the methods described in the main paper with the new evaluation metric varying the number of annotations per instance. As we would expect the results increase, given that the chance to match a given instance increase by k. Figure 1 gives more hints about the reasons that yields an improvement in performance by using more annotations, in this case for the SC method. As we can observe in the false positive profiles and the table next to it, the number of true positives (TP) increases by 1.3%. We consider that this improvement contributes to the mAP because it is significant in the top $1G$ bucket. Similarly, as we would expect the errors related to not fulfilling the minimum tIOU threshold (localization, confusion ang backgroud) decrease as the chance to meet it increase by the number of segments added. Note as well that the reduction of these kind of errors ties up with the increase of errors on predictions fulfilling the minimum threshold (double detection and wrong label). These cases increase given that localization and confusion errors have more chance to exceed the threshold. Finally, the rightmost side of Figure 1 shows the sensitivity of SC under the new metric. In comparison with the previous analysis (refer to Figure 4), we observe a considerable reduction in the impact over all the characteristics. Notably, the shape of the overall picture remains the same with the temporal context and length being the most promising areas to increase performance. However, this time the gap with the temporal agreement and number of instances is not as significant as before.

This metric is a step forward concerning the use of multiple annotations during testing due to disagreement. Here, we have discusses its impact on the false positive profile and sensitivity as a way to qualify its use. In future work, it would be pertinent to train the models with such metric to have a throughout perspective of its impact.

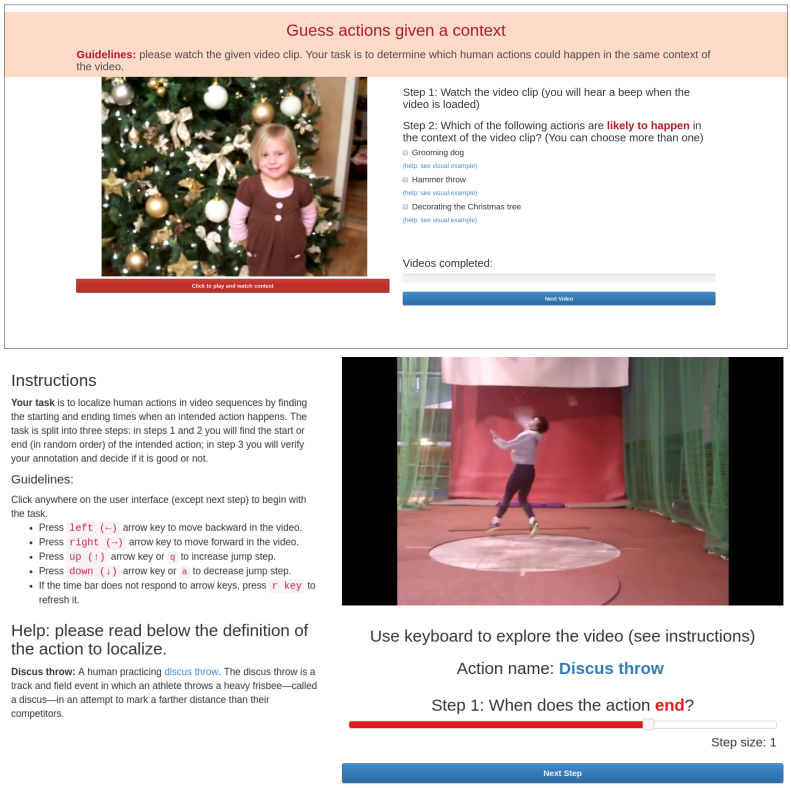

Fig. 2: **Top:** User Interface used to collect the temporal context characteristic data. **Bottom:** User Interface used to collect the temporal agreement data.

2 Additional ActivityNet Analysis

Online User Studies Interfaces. Here, we present the user interfaces used in the two online user studies introduces in the main manuscript. Figure 2 (**Top**) shows the online interface used to conduct the User Study I. We show the participant a video clip of five seconds. His/her task is to guess the action that could happen in the context of the displayed video clip. Our interface verifies that the user watches all the five seconds clip. Figure 2 (**Bottom**) illustrates the user interface used to collect the User Study II experimental data. The task of the participant consists of delimiting the starting and ending times of an action instance. An initial test of the experiment revealed that people prefer to find the start and end of the action, in that respective order, more frequently than picking the end and the start afterward. To mitigate the inherent human factor of the experiment, we ask the users to annotate each boundary at a time, and randomized the order of the question to break the unconscious human asymmetry. Additionally, we include descriptions of the actions to avoid confusions due to unknown actions.

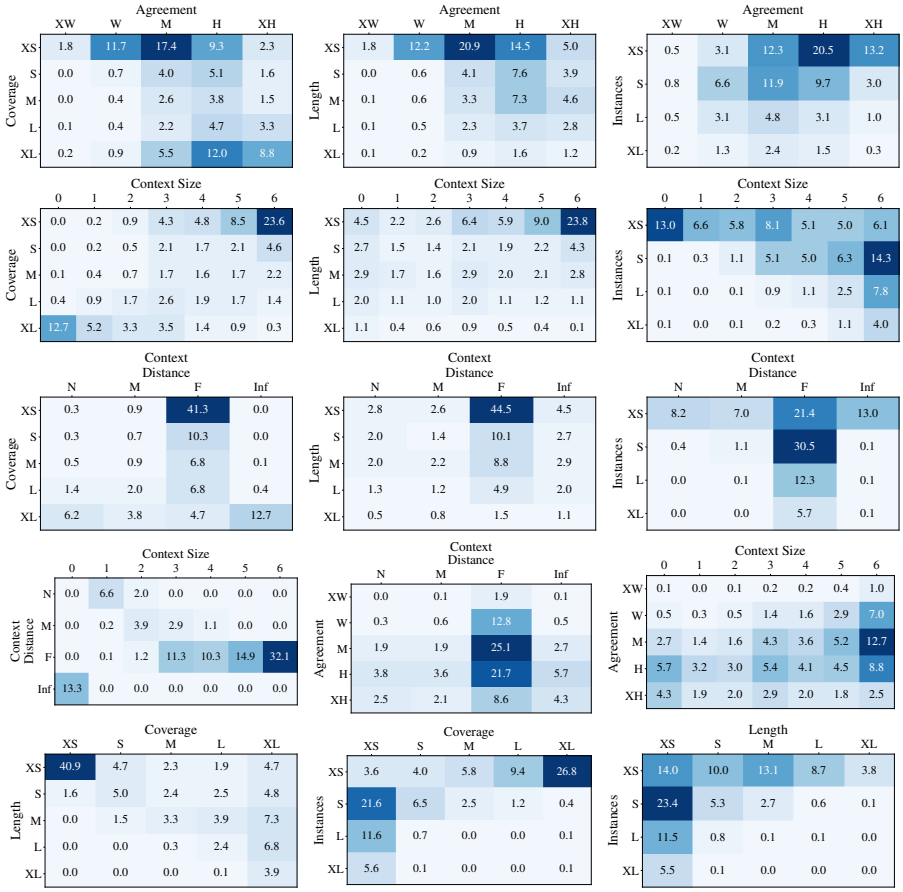

Fig. 3: We show the distribution of instances for all pairs of characteristic interactions.

Pairwise Interactions. For completeness, we show all pairwise interactions between action characteristics in Figure 3.

Average-mAP$_N$ Sensitivity. Figure 4 shows the detailed sensitivity for all the methods to action characteristics. The dashed line is the overall performance. Each bar measures the average-mAP$_N$ on a subset of ActivityNet for which a particular action characteristic holds. The figure on the right corresponds to the sensitivity profile that showcases the impact and sensitivity of each action characteristics.

False Negative Analysis. Figure 5 shows the false negative rate for all the methods on each action characteristic, and the false positive rate among three pairwise interactions.

Appropriately Reading Differences in Impact of FP Errors. Figure 5 **(Top)**, in the main paper, shows significantly different False Positive Profiles among the algorithms. For example, CES and IC have a large portion of back-

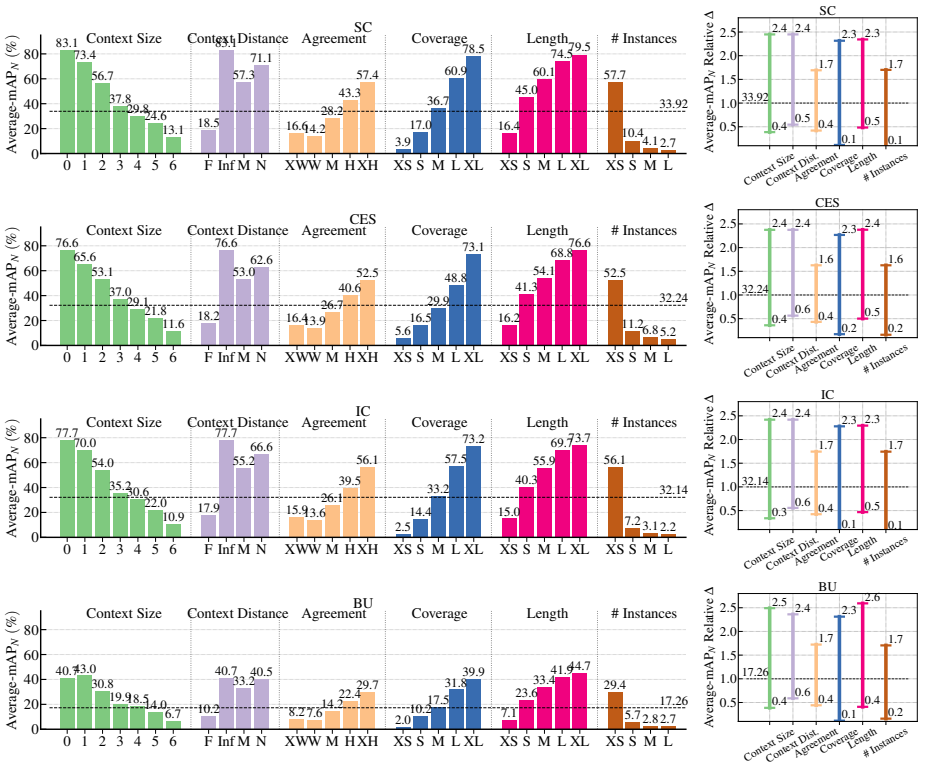

Fig. 4: Each row represents a single method. **Left:** The detailed sensitivity each method's average-mAP$_N$ to action characteristics. The dashed line is the overall performance. Each bar measures the average-mAP$_N$ on a subset of ActivityNet for which a particular action characteristic holds. **Right:** The sensitivity profile summarizing the left figure. The difference between the max and min average-mAP$_N$ represents the sensitivity while the difference between the max and the overall average-mAP$_N$ denotes the impact of the characteristic.

ground error and BU has a large portion on confusion error. However it is somehow surprising that these trends are not evident in Figure 5 **(Bottom)** where all methods have large localization error impact, while the impacts of background and confusion errors are very small. How can this be possible? In this case, everything boils down to the ranking of the predicted segments causing the error type. For example, the localization error has the most impact on average-mAP$_N$ because it is the most abundant error coming from high scoring predictions. Error from highly ranked predictions affect the area under the PR curve more than errors from low ranking predictions. To demonstrate this, Figure 6 shows the same analysis of Figure 5 **(Bottom)** using only the top-1G predictions. We observed a similar impact pattern when only considering the top-1G predictions. In other words, localization error is the most impactful error type.

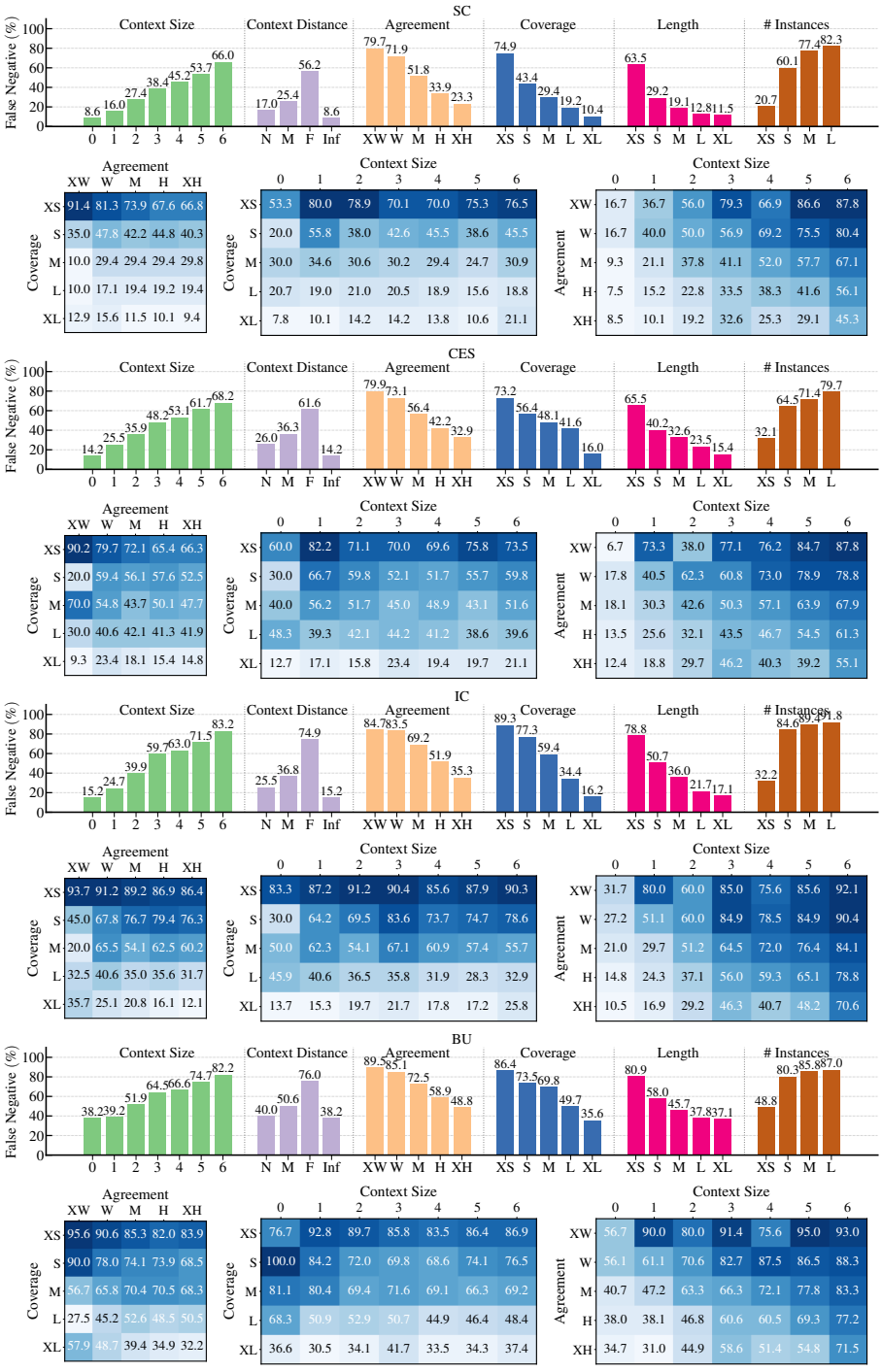

Fig. 5: Each set of bar plot and three matrices represents one method. False negative rate for each characteristic **(Top)** and three pairs of characteristics **(Bottom)**

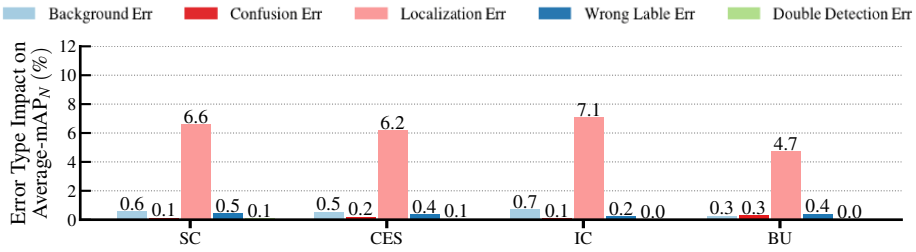

Fig. 6: The impact of error types on the average-mAP$_N$ for the top-1G predictions. In comparison with the Figure 5 in the main paper, we observe the same impact pattern for the top-1G predictions for all the methods. This reinforces the idea that the localization error is the most notable source of error and explains why only fixing the other types of errors will not yield a more significant impact.

3 THUMOS14 Analysis

In this section, we exemplify our diagnostic tool in the THUMOS14 dataset [4].

Evaluation Framework. In contrast with ActivityNet, THUMOS14 is a sports centric dataset with only 20 action categories. The dataset comprises 2584 untrimmed videos, 1010 labeled as validation set and 1574 as test set. Additionally, it provides 13320 trimmed videos labeled as training set and 2500 background videos. Typically, researchers train their models in the validation set and report results directly on the public test set. Indeed, it is a common practice to only report results on a subset of 213 videos of the original testing set which have temporal annotations. Similarly, validation videos without annotations are also discarded. Following the standard practice used by the community, we only report results for a tIoU threshold equal to 0.5.

Algorithms. For this pilot experiment, we consider the top three ranked algorithms with best performance. Table 2 summarizes the results of all the approaches considered in this study. All the methods tackled the problem in a two-stage fashion, using a proposal algorithm [7,2] followed by a classification scheme [8,9,10]. However, there are subtle design differences which are relevant to highlight.

SSN [13]. This work is the state-of-the-art peer-reviewed approach on THUMOS14 at the time of the submission. It employs a temporal grouping heuristic for generating actions proposals [11] from dense actioness predictions. The proposals are classified and refined in a subsequent stage by another sub-network. This sub-network applies a temporal pyramid pooling around the region spanned by a proposal segment, and combines the information inside the segment and the context information around it.

R-C3D [12]. It corresponds to the runner-up peer-reviewed approach in THUMOS14 at the time of the submission. Inspired by the Faster RCNN architecture [6], this framework introduces a temporal proposals network to generate candidate temporal segments of varied lengths. These segments are classified and refined by a multi-layer fully connected network in a subsequent stage. In comparison with SSN [13], that exploits optical flow as input cue to the network, this

Table 2: Localization performance as measured by mAP and mAP_N at 0.5 tIoU on THUMOS14. We show the two metrics for all predictions and for the top-10G predictions, where G is the number of ground truth instances. Using mAP_N gives slightly higher values. Notably, limiting the number of predictions to the top-10G gives performance values similar to those when considering all predictions.

	mAP (%)		**mAP**$_N$(%)	
Method	All	top-10G	All	top-10G
CMS-RC3D	40.04	39.70	43.67	43.03
SSN	29.32	29.32	31.14	31.14
R-C3D	29.89	29.89	32.38	32.38

work only relies on the RGB stream and learns a motion representation with 3D convolutions pre-trained on the large Sports-1M dataset [5].

Despite of the fact of being the runner-up peer-reviewed method by ICCV 2017, the authors provided us predictions that yield better results than SSN. We trust on the good will of the authors and limit our study to their predictions. It is worth to note that detecting the cause of the difference in the results is out of the scope of our study, and we will provide the software tools to ammend the results, if needed.

CMS-RC3D [1]. This approach corresponds to the new non-peer-reviewed state of the art in THUMOS14. This works extends the R-C3D framework to work deal with different temporal scales more effectively. Interestingly, this work improves upon the previous register mAP by 12% and claims to be designed to tackle the inherent temporal variability of the actions. These characteristics makes of it an interesting case for our insightful diagnostic.

Dataset Characterization. We provide three action characteristics for THUMOS14 *Length.* We measure length as the instance duration in seconds. We create five different length groups: Extra Small (XS: $(0, 3]$), Small (S: $(3, 6]$), Medium (M: $(6, 12]$), Large (L: $(12, 18]$), and Extra Large (XL: > 18). *Coverage.* To measure coverage, we normalize the length of the instance by the duration of the video. We categorize coverage values into five buckets: Extra Small (XS: $(0, 0.02]$), Small (S: $(0.02, 0.04]$), Medium (M: $(0.04, 0.06]$), Large (L: $(0.06, 0.08]$), and Extra Large (XL: > 0.08,). *Number of Instances.* We assign each instance the total count of instances in its video. We create four categories for the number of instances (# Instances) characteristic: Extra Small (XS: 1); Small (S: $[2, 40]$); Medium (M: $(40, 80]$); Large (L: > 80). Figure 7 shows the distribution of action characteristics in THUMOS14 as well as their pairwise interactions.

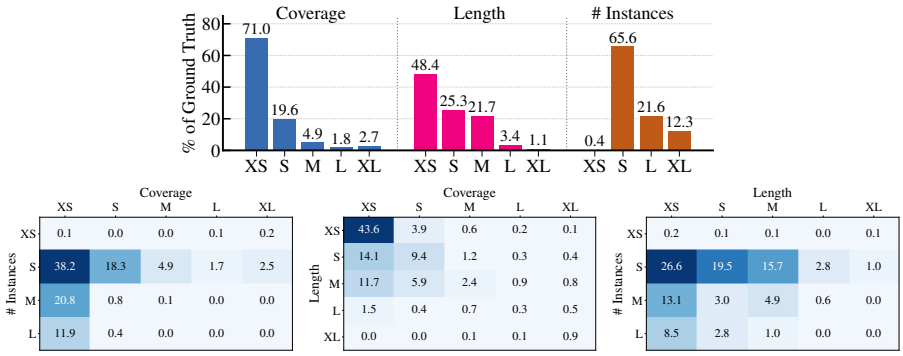

Fig. 7: Each set of bar plot and three matrices represents one method. False negative rate for each characteristic **(Top)** and the three pairs of characteristics **(Bottom)**.

False Positive Analysis. Figure 8 shows the false positive profiles of the three THUMOS14 methods. Each profile demonstrates the FP error breakdown in the

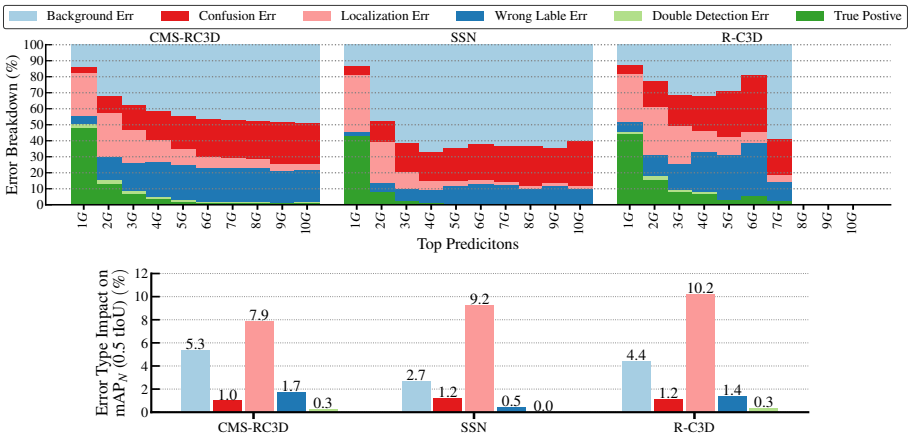

Fig. 8: **Top:** The false positive profiles of the three THUMOS14 methods. Each profile demonstrates the FP error breakdown in the top-10G predictions. **Bottom:** The impact of error types on the mAP$_N$ (0.5 tIoU), *i.e.* the improvement gained from removing all predictions that causes each type of error. The Localization Error (pink bar) has the most impact.

top-10G predictions. It also highlights the impact of error types on the mAP$_N$ (0.5 tIoU), *i.e.* the improvement gained from removing all predictions that cause each type of error.

Average-mAP$_N$ Sensitivity. Figure 9 shows the detailed sensitivity of each method mAP$_N$ (0.5 tIoU) to action characteristics. The dashed line represents the overall performance. The sensitivity profile in the rightmost side of the figure showcases the overall sensitivity for each action characteristic as well as its impact.

False Negative Analysis. Figure 10 illustrates the false negative rate for each action characteristic and the interactions between the three pairs of characteristics.

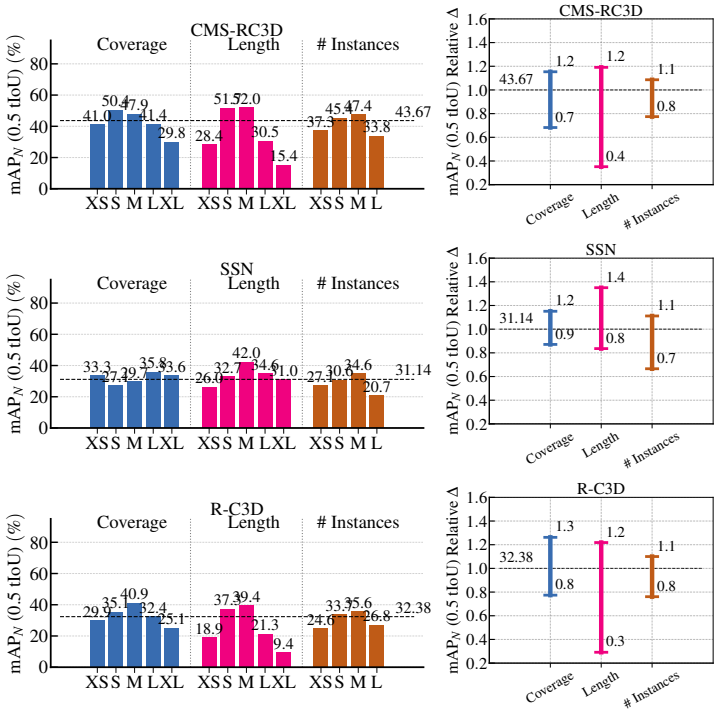

Fig. 9: Each row represents a single method. **Left:** The detailed sensitivity each method's mAP$_N$ (0.5 tIoU) to action characteristics. The dashed line is the overall performance. Each bar measures the mAP$_N$ (0.5 tIoU) on a subset of THUMOS14 for which a particular action characteristic holds. **Right:** The sensitivity profile summarizing the left figure. The difference between the max and min mAP$_N$ represents the sensitivity while the difference between the max and the overall mAP$_N$ denotes the impact of the characteristic.

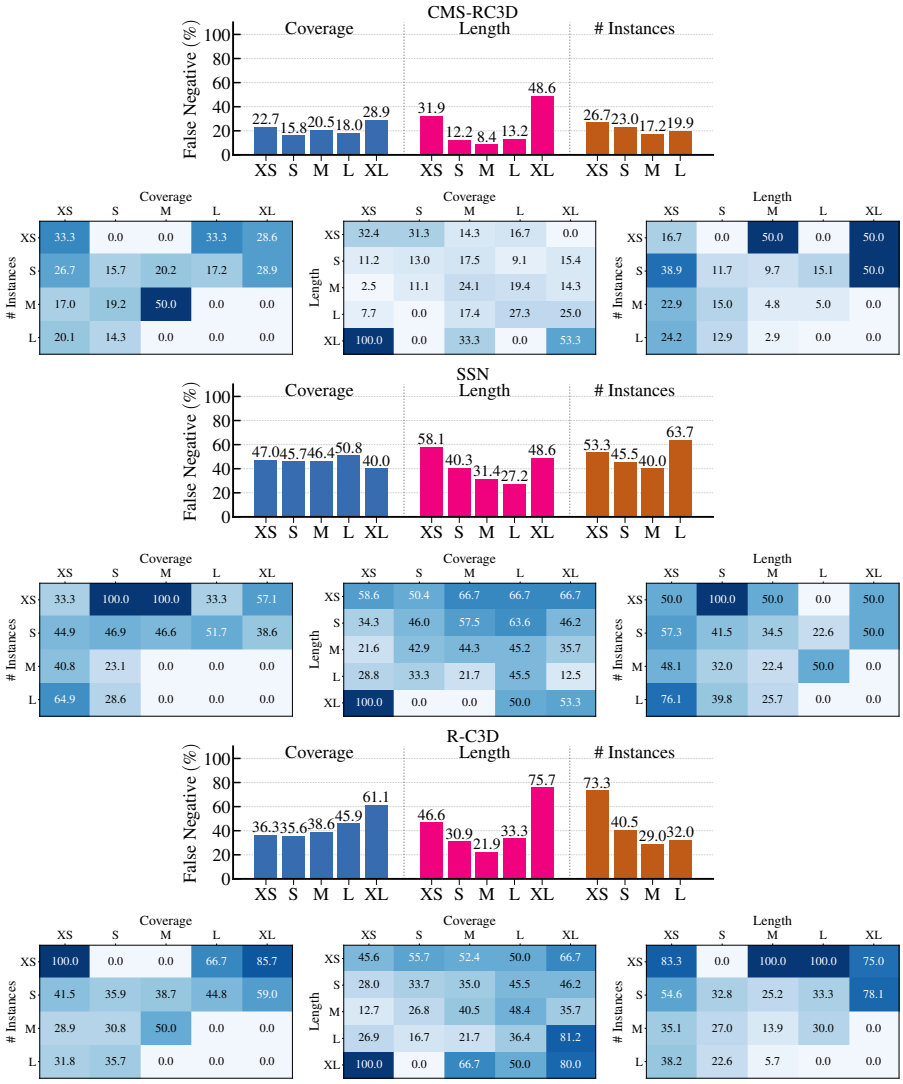

Fig. 10: Each set of bar plot and three matrices represents one method. False negative rate for each characteristic **(Top)** and the three pairs of characteristics **(Bottom)**.

References

1. Bai, Y., Saenko, H.X.K., Ghanem, B.: Contextual multi-scale region convolutional 3d network for activity detection. CoRR (2017)
2. Gao, J., Yang, Z., Sun, C., Chen, K., Nevatia, R.: Turn tap: Temporal unit regression network for temporal action proposals. In: ICCV (2017)
3. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR 2015. pp. 961–970 (2015). https://doi.org/10.1109/CVPR.2015.7298698, https://doi.org/10.1109/CVPR.2015.7298698
4. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/ (2014)
5. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
7. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: CVPR (2016)
8. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
9. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
10. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Val Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV (2016)
11. Xiong, Y., Zhao, Y., Wang, L., Lin, D., Tang, X.: A pursuit of temporal accuracy in general activity detection. CoRR **abs/1703.02716** (2017), http://arxiv.org/abs/1703.02716
12. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: ICCV (2017)
13. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: ICCV 2017 (Oct 2017)