Supplementary Material
PivoTAL: Prior-Driven Supervision for Weakly-Supervised Temporal Action Localization

Mamshad Nayeem Rizve*‡ Gaurav Mittal*† Ye Yu† Matthew Hall† Sandra Sajeev†
Mubarak Shah‡ Mei Chen†
†Microsoft ‡University of Central Florida
{gaurav.mittal, yu.ye, mathall, ssajeev, mei.chen}@microsoft.com
nayeemrizve@knights.ucf.edu shah@crcv.ucf.edu

In this document, we provide additional analysis of our method PivoTAL, both quantitatively and qualitatively. To be particular, we provide additional ablation experiments (top-down) in Section 1. Next, we provide an analysis of the effect of different β values for Gaussian mask generation in Section 2. After that, in Section 3, we provide a detailed analysis to show the impact of our per-class confidence normalization. In Seciont 4, we provide details of inference pipeline for our Prior-driven Localization Head. Finally, in Section 5, we provide additional visual results demonstrating the effectiveness of PivoTAL over *Base WTAL*.

1. Additional Ablation Study

Method	AVG mAP
PivoTAL w/o confidence propagation	39.0
PivoTAL w/o per-class normalization	46.4
PivoTAL w/o supervised MIL	46.6
PivoTAL w/o scene prior	48.2
PivoTAL w/o Gaussian prior	48.3
PivoTAL w/o learning-based actionness head	48.9
PivoTAL	**49.6**

Table 1. Ablation studies on the **THUMOS-14** dataset showing the effectiveness of each component of PivoTAL.

We report results for additional top-down ablation experiments in Table 1. The first row shows that if we only rely on SG to generate the pseudo-GT action snippets without utilizing the confidence of the predictions from the MIL-based WTAL head, the performance deteriorates by more than 10%. The next row demonstrates the effectiveness of per-class confidence normalization. We observe that the performance goes down by 3.2% if no normalization is performed. This result validates our hypothesis that the per-class normalization re-normalizes the confidence of relatively harder or under-represented classes, thereby boosting the overall performance. The third row demonstrates the impact of MIL loss on the prior-driven localization head, which penalizes the errors in video-level classification scores, and removing it lowers the performance by 3%. After that, we observe the effectiveness of our scene prior injection in improving the action boundaries. We notice that, even though this particular prior only works in improving the action boundaries, it influences the overall performance by a noticeable margin (1.4%). The results reported in Row 5 shows that the performance drops by 1.3% without the Gaussian prior, validating the effectiveness of locally consistent actionness scores. Finally, the second to last row shows the performance goes down by 0.7% without the learning-based actionness head. The drop is not large, showing that our Gaussian-based local smoothness prior alone is effective in generating locally consistent foreground actionness scores. At the same time, this validates that $\mathbf{a}_{fg}$ is still required to achieve the best performance as it can complement $\mathbf{a}_{gauss}$.

2. Analysis of β

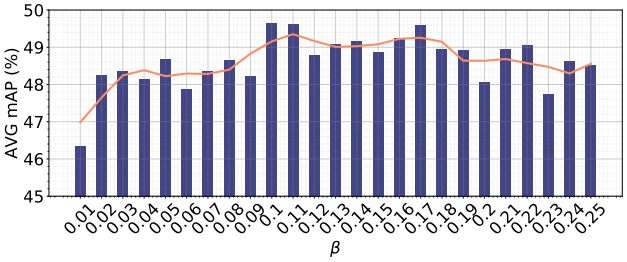

Figure 1. Effect of changing β values for the Gaussian prior on the **THUMOS-14** dataset. We observe that the performance predictably deteriorates at large and small β values.

*Authors with equal contribution.
This work was done as Mamshad's internship project at Microsoft.

We use a hyperparameter β to control the variance of the generated Gaussian mask, $\mathbf{G}$, for injecting learnable Gaussian priors into actionness score generation. For all of our experiments, we set the value of β to 0.1. Here, we conduct additional experiments on the THUMOS-14 dataset by varying the value of β. We report the results in Figure 1. We observe that PivoTAL performs relatively stable over β values in the range $0.10 - 0.17$. We observe that smaller and higher values deteriorate performance. This is to be expected since a smaller value will increase the variance leading to a constant mask over the entire video and relatively higher β will create very small localized masks. In both of these cases, the resultant Gaussian mask will be ineffective in generating representative actionness scores leading to poor performance.

3. Effect of Normalization

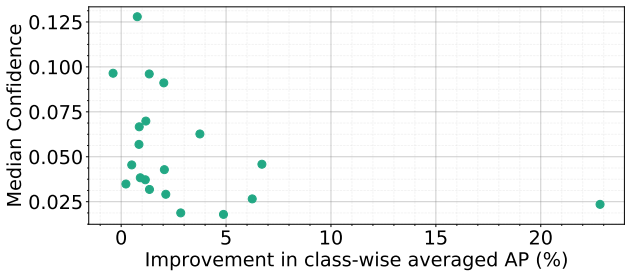

Figure 2. Effect of per-class confidence normalization on the **THUMOS-14** dataset. We observe that the performance of classes with lower median confidence scores shows relatively higher improvement.

To train the *Prior-driven Localization Head* in a *localization-by-localization* manner, we perform confidence-aware self-training based on the pseudo-action snippets, $\mathbb{A}$, which takes advantage of both weak annotations and human priors. While doing so, to address the issue of overconfident predictions on the easier classes and to improve the performance of the relatively harder and underrepresented classes, we perform per-class predictive confidence normalization. Our ablation studies (Table 3 in the main text) demonstrate that removing this per-class normalization deteriorates the average mAP by 3% on the THUMOS-14 dataset. To further investigate the effect of per-class normalization, we perform a per-class performance analysis. We report the results in Figure 2 with x-axis denoting the class-wise averaged AP (%) and y-axis denoting the median confidence score per class. We observe that generally, the classes with lower median confidence scores have higher improvement. To be particular, we observe that the average AP of the class *Cliff Diving* improves by 22.8% which has one of the lowest median confidence scores, 0.023 (with confidence scores ranging from 0.018 to 0.128 over all classes). The same trend is observed for the action *Long Jump* where the performance improved by 6.2%, which has a very low median confidence score of 0.026. This empirically validates that our per-class confidence normalization improves the performance of relatively harder (having lower overall confidence) classes.

4. Inference Pipeline

During inference, the *Base WTAL Head* is not used. We directly feed clip-level video features to *Prior-driven Localization Head* and obtain the output from its classifier and regression heads. We then subject this output to non-maximum suppression (NMS) to remove overlapping predictions and obtain the final action snippets.

5. Visualizations

In this section, we provide additional visualizations comparing the performance of PivoTAL with *Base WTAL*. The visualization are provided in Figure 3, 4, 5, 6, and 7. In the figures, Green denotes ground truth action snippets, Purple denotes true positive action snippets, Pink denotes false positive action snippets. For these visualizations, we consider an IoU overlap of at least 0.1 to be true positive and below 0.1 to be false positive. We observe that, unlike *Base WTAL*, in most cases, the PivoTAL can detect all the ground-truth instances correctly. Besides, even in the cases when *Base WTAL* detects a ground-truth instance correctly, we find the PivoTAL's detection to be better aligned and more complete with respect to the ground-truth.

For instance, in Figure 3 (Top) we observe that *Base WTAL* misses the *Javelin Throw* action around 1025 seconds, whereas the same action is detected by PivoTAL while covering almost the entire temporal extent. A similar trend is observed in Figure 3 (Bottom), where *Base WTAL* misses the *Hammer Throw* action (correctly detected by PivoTAL) around 110 seconds. In Figure 4 (Top), we see another instance where *Base WTAL* misses the *Javelin Throw* action around 500 seconds. Figure 4 (Bottom) demonstrates a similar case for the *Pole Vault* action around 860 seconds. The 625 seconds of Figure 5 (Top) and 60 seconds of Figure 5 (Bottom) show similar instances for *Javelin Throw* and *Hammer Throw* actions respectively where PivoTAL can correctly detect a ground-truth instance that is missed by *Base WTAL*.

Figure 6 (Top) shows an interesting example where both *Base WTAL* and PivoTAL detect all the ground-truth instances correctly for the *Shotput* action. However, upon closer inspection, it is evident that the detections from the PivoTAL have better temporal overlap with the ground-truth instances. One thing to note here is that the performance of the PivoTAL suffers in the initial part of the video. We hypothesize that this happens because the size of the fore-

ground is considerably small compared to the background and the foreground also suffers from poor background separation. We believe this can be somewhat alleviated by using stronger backbones that can deal with higher input resolutions. Figure 6 (Bottom) demonstrates another example where *Base WTAL* misses an action around 1025 seconds which is correctly detected by PivoTAL. Finally, Figure 7 demonstrates examples where both *Base WTAL* and PivoTAL detect all ground-truth instances but the predictions from PivoTAL are better aligned with the ground-truth. For instance, the detection from PivoTAL around 225 seconds on Figure 7 (Top) has better alignment with the ground-truth than the one from *Base WTAL*. We can draw a similar conclusion by looking at around 130 seconds of Figure 7.

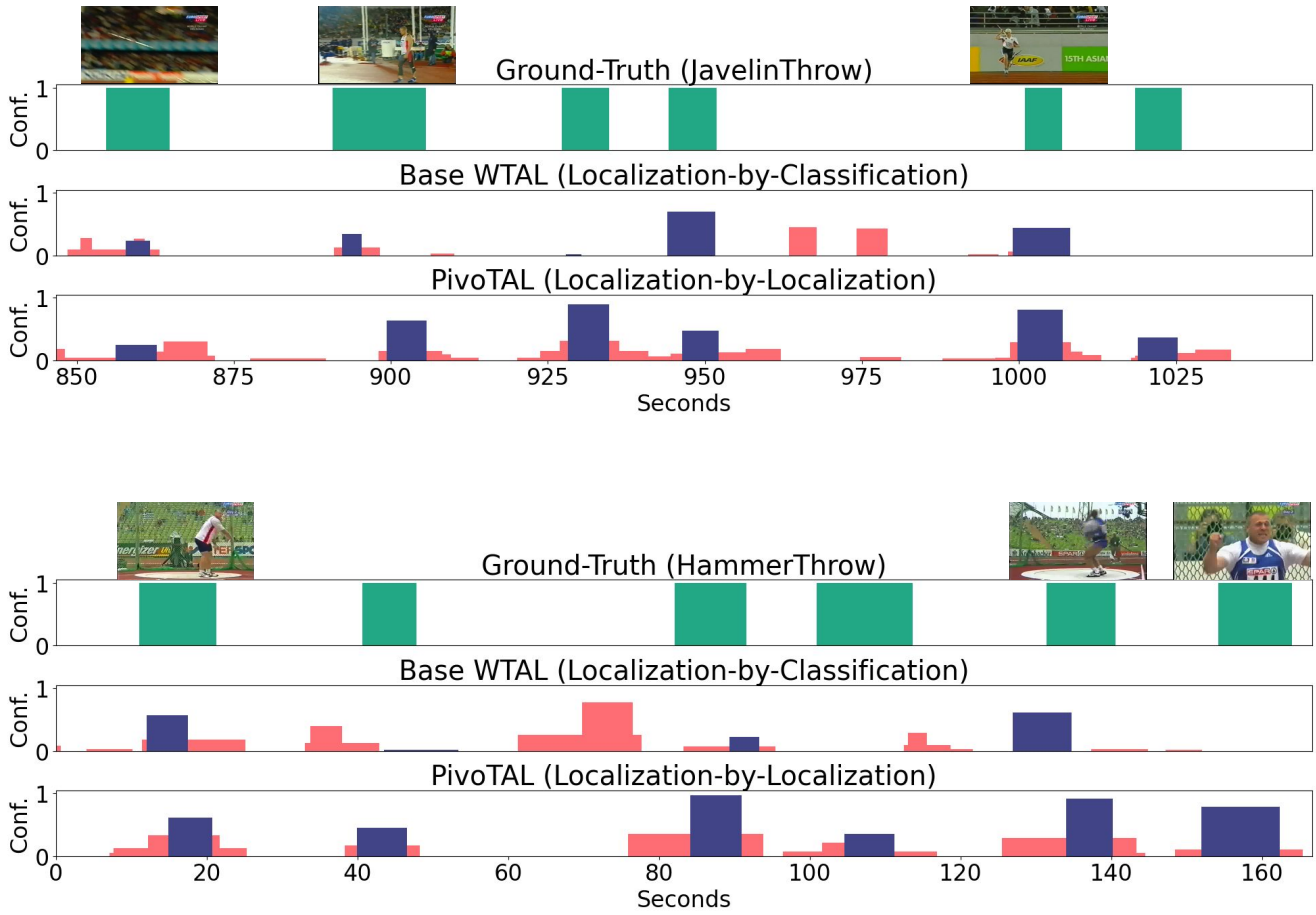

Figure 3. Visualizations of *Base WTAL* and PivoTAL's predictions. Green denotes ground truth action snippets, Purple denotes true positive action snippets, Pink denotes false positive action snippets. We observe that unlike *Base WTA*L, PivoTAL detects all the ground-truth instances. We also notice that PivoTAL's predictions are better aligned with the ground-truth. In the top figure, we observe that the *Base WTAL* misses the *Javelin Throw* action around 1025 seconds, whereas the same action is detected by PivoTAL covering almost the entire temporal extent. A similar trend is observed in the bottom figure, where the *Base WTAL* misses the *Hammer Throw* action (correctly detected by PivoTAL) around 110 seconds.

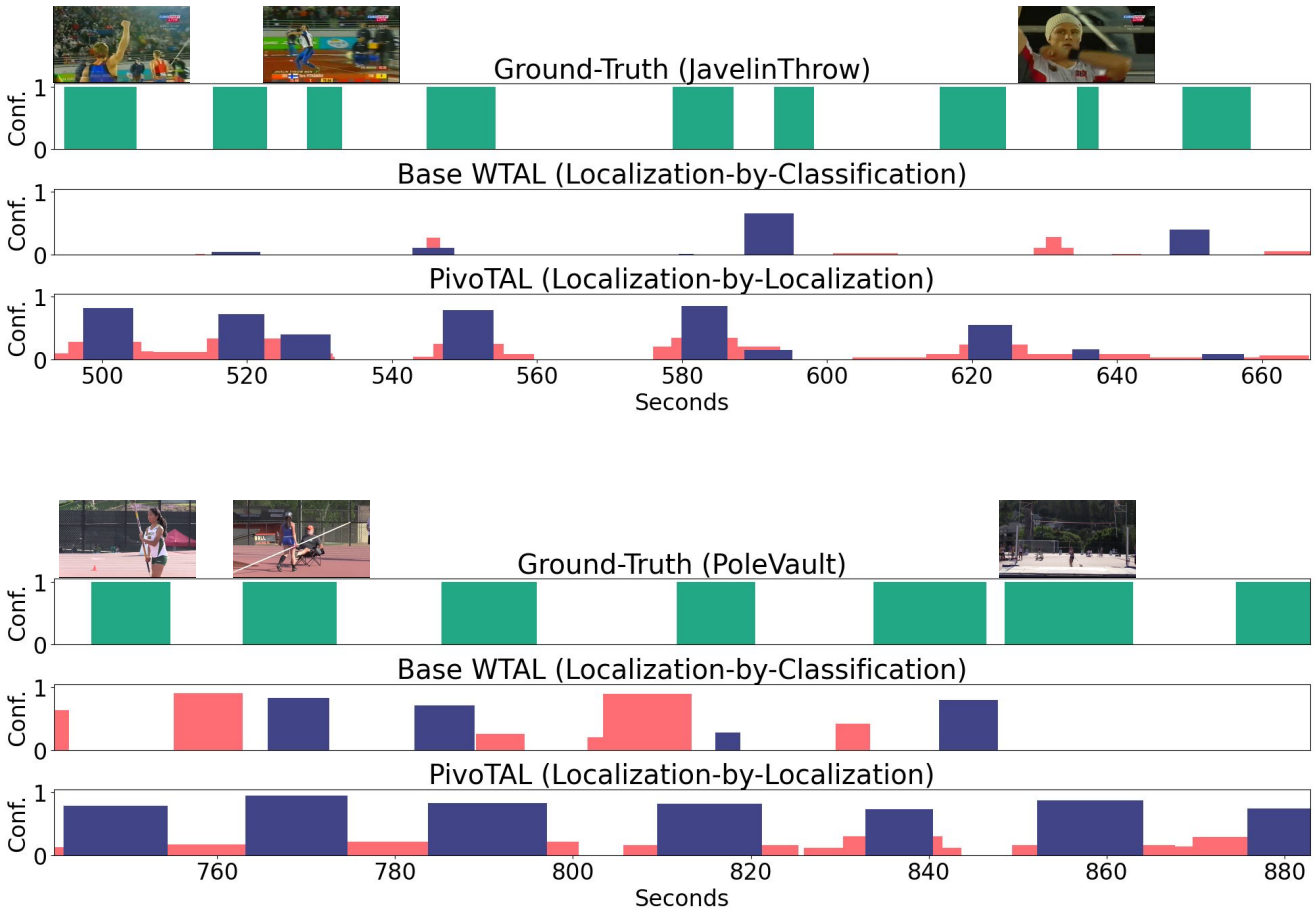

Figure 4. Visualizations of *Base WTAL* and PivoTAL's predictions. Green denotes ground truth action snippets, Purple denotes true positive action snippets, Pink denotes false positive action snippets. We observe that, unlike *Base WTAL*, PivoTAL detects all the ground-truth instances. We also notice that PivoTAL's predictions are better aligned with the ground-truth. In the top figure, we observe that the *Base WTAL* misses the *Javelin Throw* action around 500 seconds, whereas the same action is detected by PivoTAL. A similar trend is observed in the bottom figure, where the *Base WTAL* misses the *Pole Vault* action (correctly detected by PivoTAL) around 860 seconds.

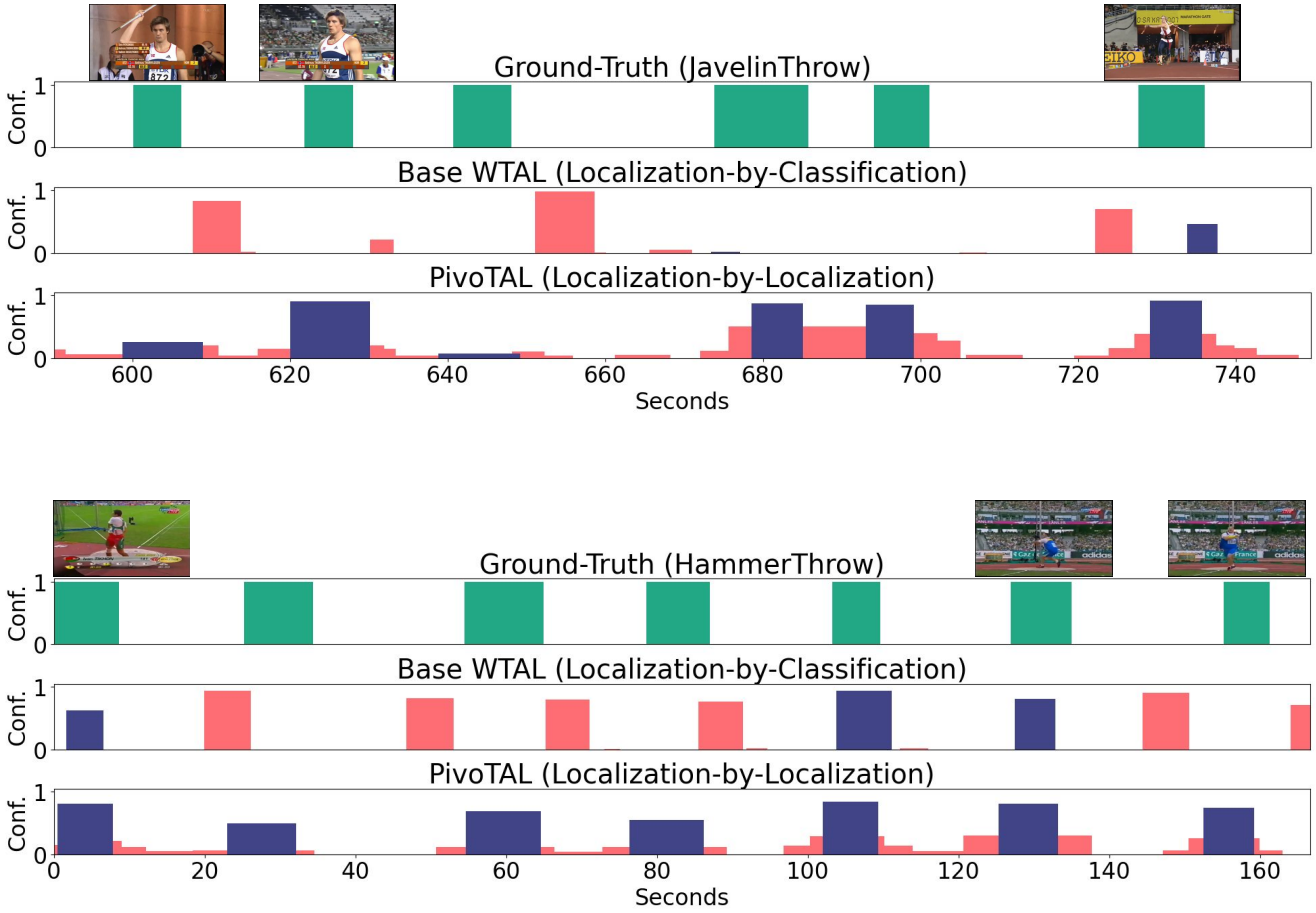

Figure 5. Visualizations of *Base WTAL* and PivoTAL's predictions. Green denotes ground truth action snippets, Purple denotes true positive action snippets, Pink denotes false positive action snippets. We observe that, unlike *Base WTAL*, PivoTAL detects all the ground-truth instances. We also notice that PivoTAL's predictions are better aligned with the ground-truth. In the top figure, we observe that the *Base WTAL* misses the *Javelin Throw* action around 625 seconds, whereas the same action is detected by PivoTAL. A similar trend is observed in the bottom figure, where the *Base WTAL* misses the *Hammer Throw* action (correctly detected by PivoTAL) around 60 seconds.

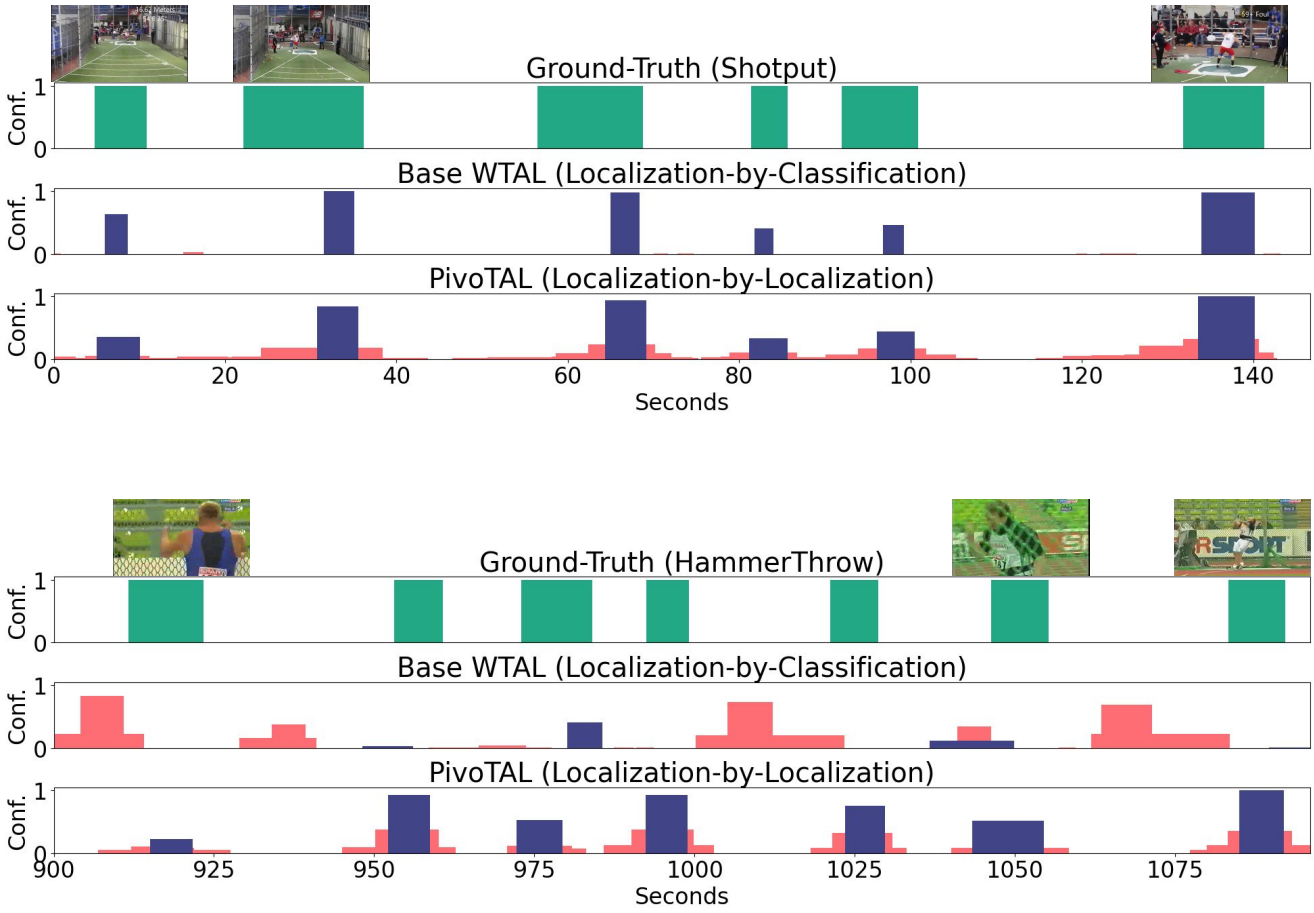

Figure 6. Visualizations of *Base WTAL* and PivoTAL's predictions. Green denotes ground truth action snippets, Purple denotes true positive action snippets, Pink denotes false positive action snippets. We observe that, unlike *Base WTAL*, PivoTAL detects all the ground-truth instances. We also notice that PivoTAL's predictions are better aligned with the ground-truth. In the top figure, we observe that both *Base WTAL* and PivoTAL detect all the ground-truth instances correctly for the *Shotput* action. However, upon closer inspection, it is evident that the detections from the PivoTAL have better temporal overlap with the ground-truth instances. In the bottom figure, we observe that the *Base WTAL* misses the *Hammer Throw* action around 1025 seconds, whereas the same action is detected by PivoTAL.

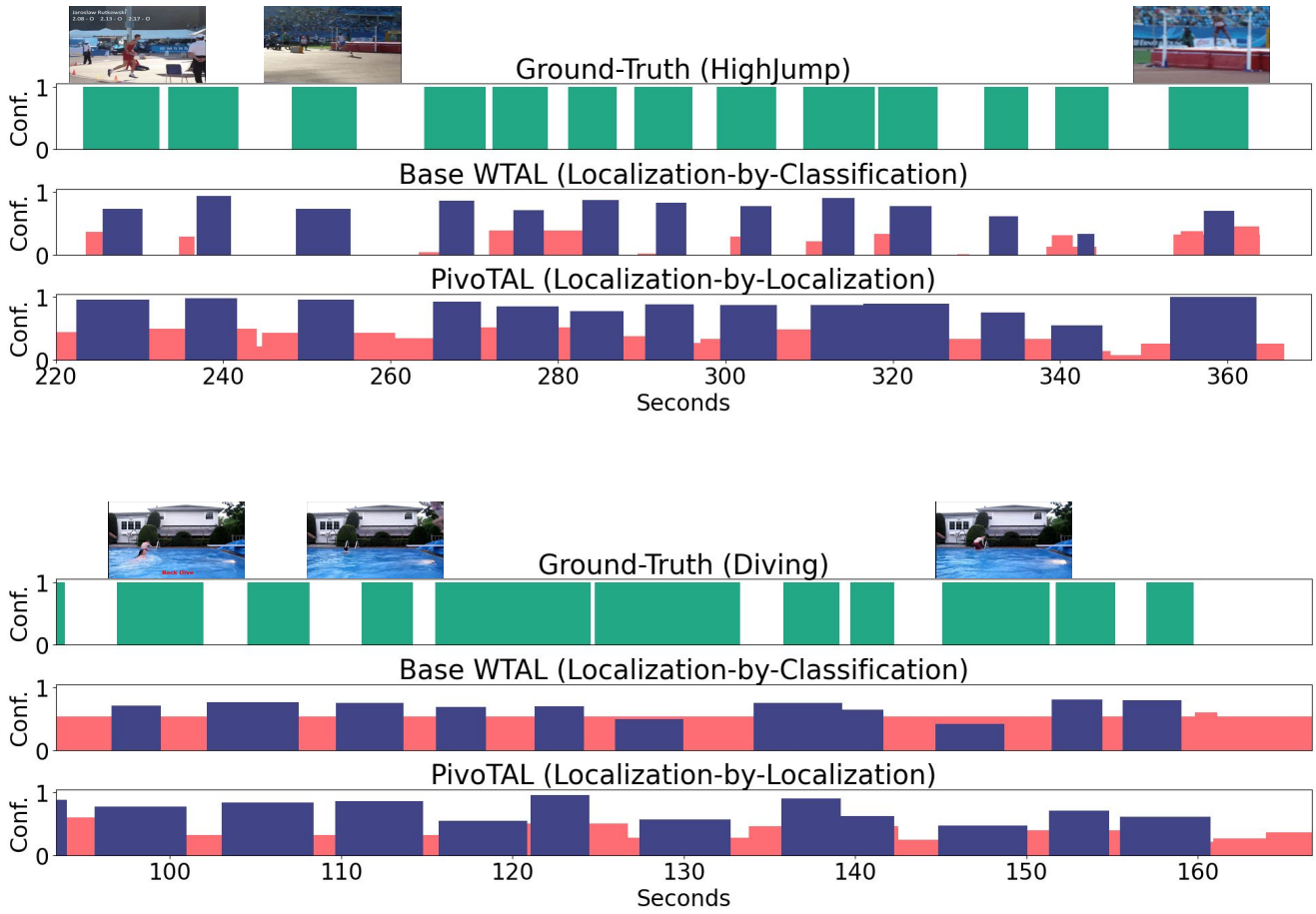

Figure 7. Visualizations of *Base WTAL* and PivoTAL's predictions. Green denotes ground truth action snippets, Purple denotes true positive action snippets, Pink denotes false positive action snippets. We observe that even when *Base WTAL* detects all the ground-truth instances correctly, PivoTAL's predictions are better aligned with the ground-truth. In the top figure, we observe that the detection from PivoTAL around 225 seconds for *High Jump* action has better alignment with the ground-truth than the one from *Base WTAL*. A similar trend is observed in the bottom figure, where the *Diving* action around 130 seconds has a relatively better alignment with the predictions from PivoTAL.