The Blessings of Unlabeled Background in Untrimmed Videos (Supplementary Material)

This supplementary material includes additional experimental results that are not presented in the main paper and more qualitative results to demonstrate the superior performance of our proposed TS-PCA deconfounder.

1. Visualization of CAS Scores

To further demonstrate the effectiveness of the TS-PCA, we plot both the CAS generated by BasNet [3] and the CAS generated by incorporating TS-PCA (+TS-PCA) deconfounder in Figure 1, where the temporal CAS scores are normalized to 0 to 1. Compared with BasNet, the CAS generated by +TS-PCA can cover large and dense regions to obtain more accurate action segments. In the example of Figure 1, +TS-PCA can discover almost all the actions that are labeled in the ground-truth. However, BasNet tends to only detect the most salient regions of action segments. These results further demonstrate the effectiveness of the TS-PCA deconfounder in calibrating CAS. Specifically,

First Example: BasNet fails to localize CLIFFDIVING with short temporal durations while CAS calibrated by TS-PCA succeeds to find the last segment.
Second Example: BasNet has difficulty in detecting action instances with long duration perfectly, where it tends to generate incomplete instances. The calibrated CAS can tackle this problem by increasing the CAS scores of false-negative segments.
Third Example: When the background is semantically correlated with the foreground, it is difficult to well separate them. It can be observed that the calibrated CAS suppresses feedback of background to some extent.

2. More Qualitative Results

More qualitative results are illustrated in Figure 2. The first three rows are videos from the testing set of THUMOS-14 [2] and the last two rows are from the validation set of ActivityNet-1.3 [1]. It can be observed that, with the TS-PCA deconfounder to calibrate CAS, more accurate action instances are localized, which demonstrates the effectiveness of the proposed deconfounder strategy.

3. Failure Cases

We also show some failure cases in Figure 3. For action instances which only occupy small spatial regions of the whole frame (first two rows), it is challenging to well localize them. Moreover, if the videos are of low quality (last row), it will be hard to capture the corresponding semantic meanings, which thereby results in inaccurate proposals.

References

[1] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[2] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. In *ECCVW*, 2014.

[3] P. Lee, Y. Uh, and H. Byun. Background suppression network for weakly-supervised temporal action localization. In *AAAI*, 2020.

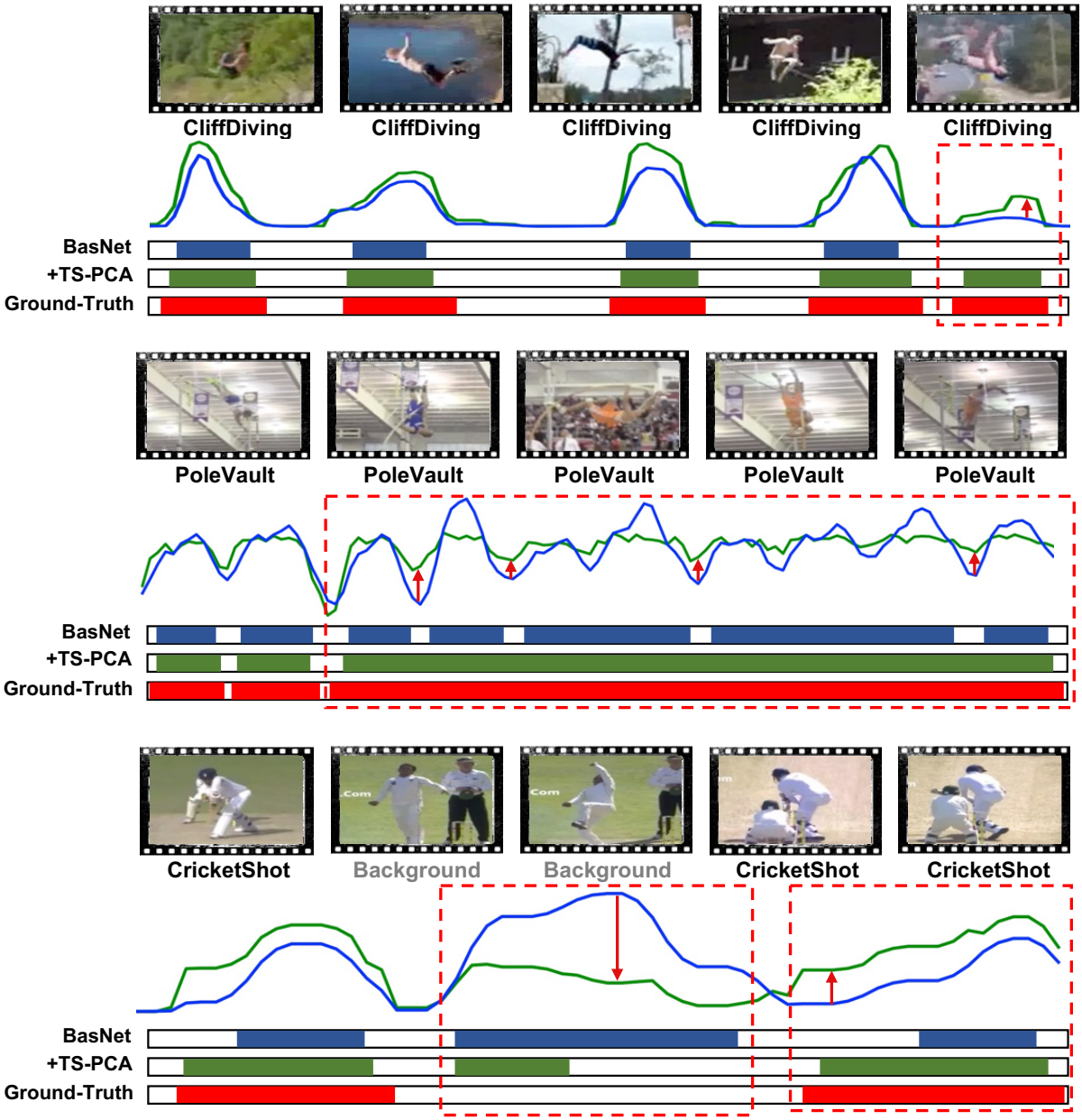

Figure 1: Comparisons of CASs generated by BasNet and TS-PCA on three examples from THUMOS-14. The blue line represents the CASs generated by BasNet and the green line represents the CASs generated by incorporating TS-PCA. Compared with BasNet, the CAS generated by TS-PCA can cover large and continuous regions to obtain more accurate action segments.

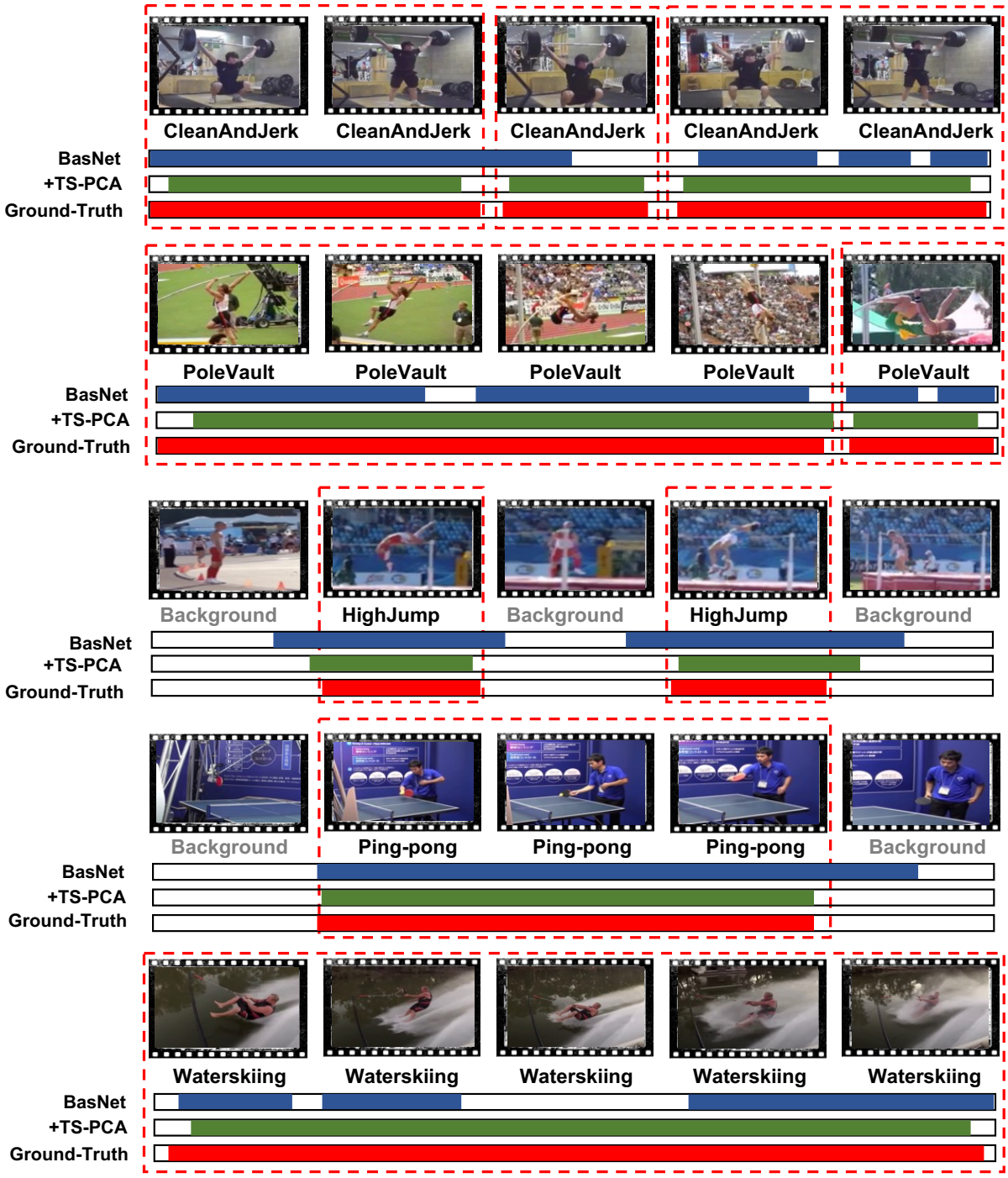

Figure 2: Qualitative results of BasNet with the TS-PCA deconfounder. First three rows show action instances on THUMOS-14. Last two rows show action instances on ActivityNet-1.3. Red dashed rectangles highlight the improved segments predicted by +TS-PCA.

Figure 3: Failure cases generated by BasNet with the TS-PCA deconfounder on THUMOS-14.