(Supplementary)
MiniROAD: Minimal RNN Framework for Online Action Detection

Joungbin An[1] Hyolim Kang[1] Su Ho Han[1] Ming-Hsuan Yang[1,2,3] Seon Joo Kim[1]
[1]Yonsei University [2]UC Merced [3]Google Research

We provide additional experimental results and analysis that support our main paper.

1. Effect of the proposed method

Apart from the primary experiments, the effect of applying our proposed method on all the datasets tested in the paper is demonstrated in Table 1. As mentioned in the main paper, *Uniform Weight* represents the traditional clip-based training approach with uniform weights for each loss at every time-step, while *Proposed* utilizes non-uniform weights proposed in our main paper. The results indicate that our proposed method consistently improves performance across all datasets compared to the conventional method. This suggests that our proposed method effectively addresses the training and inference discrepancy.

	THUMOS		TVSeries		FineAction
	Anet	Kinetics	Anet	Kinetics	Kinetics
Uniform Weight	67.5	66.4	86.0	89.0	35.7
Proposed	69.3	71.8	88.5	89.6	37.1
Whole Video	69.6	72.0	88.3	89.4	38.0

Table 1: Results of applying the proposed method over the conventional (uniform weights on the loss at each time step) method. The last row shows the result of training with the whole video. Three datasets, THUMOS'14[5], TVSeries[1], FineAction[6] are shown where Anet stands for feature extractor pretrained with Activitynet [4] and Kinetics stands for feature extractor pretrained with Kinetics [5].

2. Alleviating Training-Inference Discrepancy

In addition to the proposed method in the paper, we experiment other possible approach in alleviating the training and inference discrepancy.

Loading the Whole Video The most straightforward approach to address the training and inference discrepancy is to make the training stage identical to the inference stage, i.e., training the model with the entire video instead of using divided clips. To explore this approach, we train our model by loading one complete video at a time. As the training process now mirrors the inference process, we employ the conventional uniform weight for the loss computed at each time step. The results are presented in Table 1, where the training with the whole video is denoted by *Whole Video*. The findings reveal that using the entire video during training leads to a significant increase in performance compared to the performance of using conventional training method. Our results support the main claim of our paper that the training and inference discrepancy is the root cause of the inferior performance of RNNs and that mitigating this discrepancy is crucial for effective RNN training. However, this approach has some limitations such as the lack of support for batch training since one video must be loaded one at a time due to variable video lengths and instability in training caused by the high variance in the estimate of the gradient with small batch size [2, 3]. Therefore, we employ clip-based training with non-uniform weights proposed in our main paper, which can use batch training without requiring loading the whole video each time. Our proposed method achieves comparable performance to the training method that is free of training and inference discrepancy.

3. Qualitative Analysis

The Best and Worst Classified Actions. Figure 1 displays the top three best and worst classified actions in the THUMOS'14 dataset. It shows that the well-classified actions entail a person with significant body movements over an extended duration, while the poorly classified actions involve quick bursts of action focused on small objects.

Qualitative Comparison with SOTA We provide a qualitative comparison with the previous best-performing method, TeSTra [7], on the THUMOS'14 dataset. Based on various comparisons with TeSTra, we observed that our model and TeSTra have similar predicted confidence scores for action instances. However, the difference in perfor-

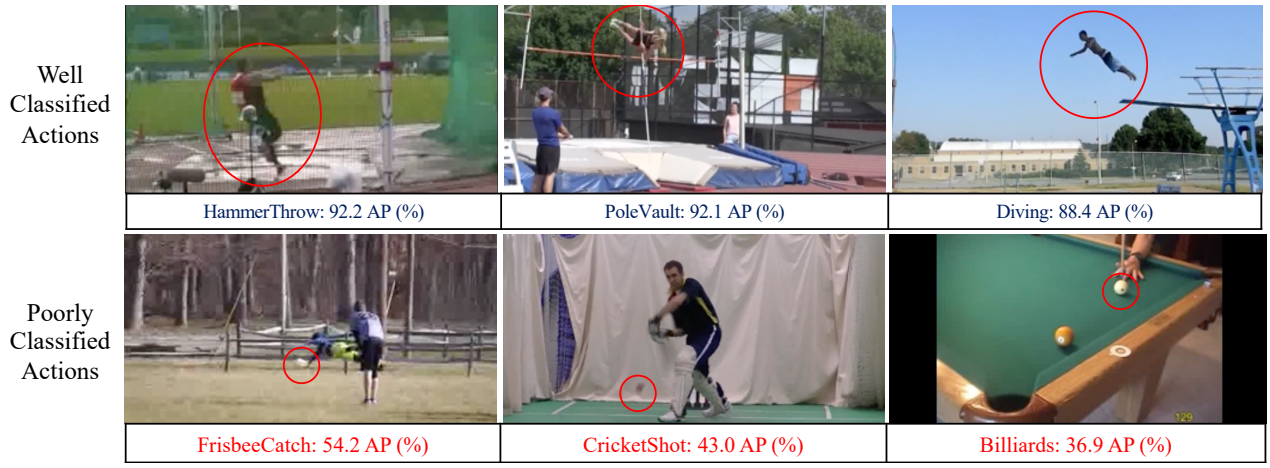

Figure 1: Class wise AP (%) of the top 3 well and poorly classified actions in THUMOS'14 dataset.

mance between TeSTra and our method stems from the number of false positives. To elaborate, we visualized the confidence scores of TeSTra and our method for four action classes that exhibit the most significant differences in performance, as shown in Figure 2. The visualized confidence scores demonstrate that our model is more effective in differentiating actions from backgrounds and is less prone to predicting false positives. Our method is more conservative in making decisions about actions, while we observed that TeSTra is more sensitive to movements such as camera and object movement.

4. Limitation and Future Work

As is common practice, most works, including our own, use motion features as input for the model. However, the computational cost of computing optical flow for these features, as analyzed in the runtime analysis section of the main paper, is high. While faster flow algorithms such as NVOFA exist, the online nature of the OAD task renders motion features unsuitable, despite their persistent use. To advance the OAD community, it is crucial to explore the potential of flow-free end-to-end methods, which have yet to be fully examined.

References

[1] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *European Conference on Computer Vision (ECCV)*, pages 269–284. Springer, 2016. 1

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 1

[3] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 1

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970. IEEE, 2015. 1

[5] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding(CVIU)*, 155:1–23, 2017. 1

[6] Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE Transactions on Image Processing*, 2022. 1

[7] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 3

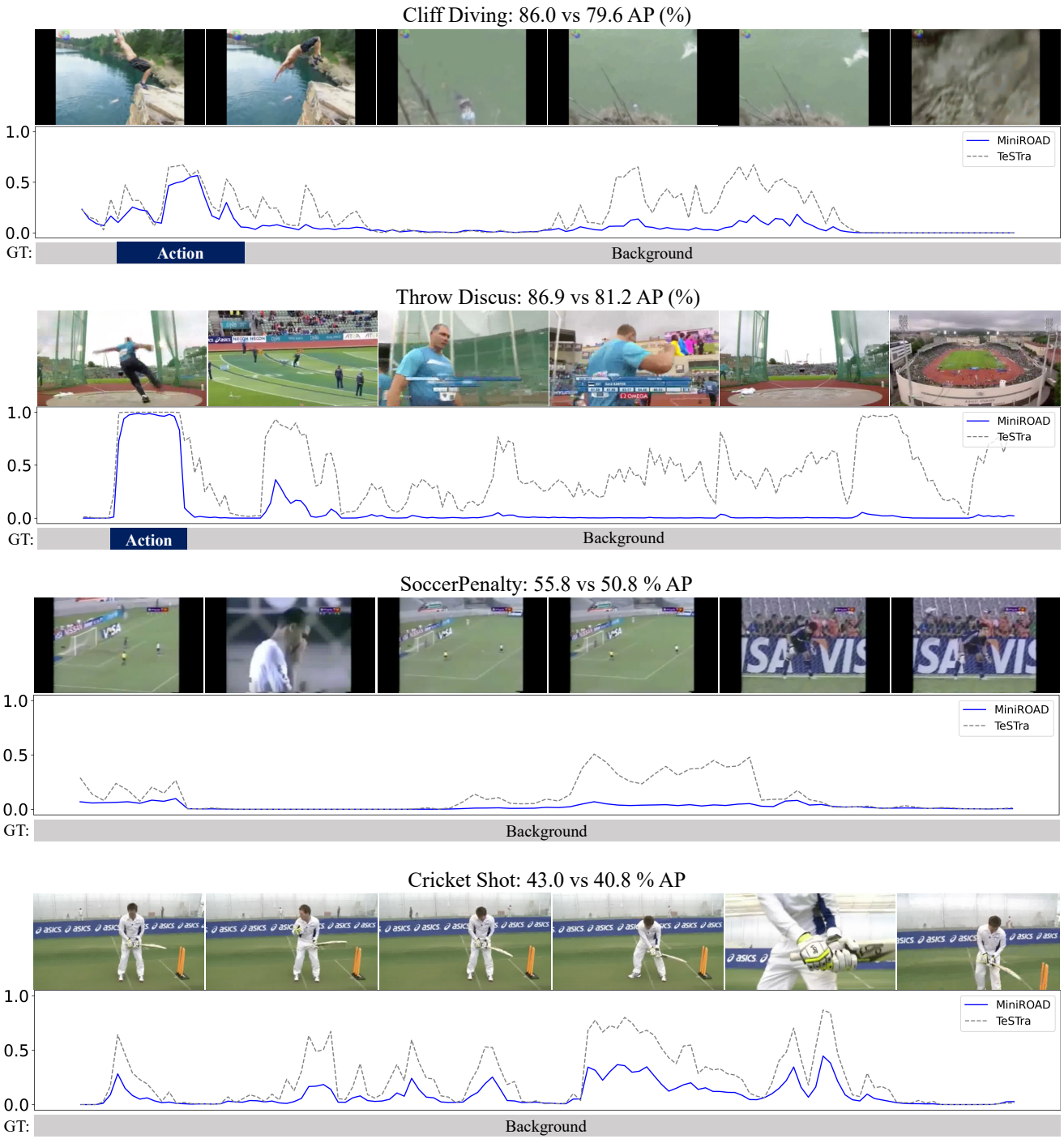

Figure 2: Qualitative comparison of ours and TeSTra [7] on THUMOS'14 on action instances that exhibit the greatest performance difference. Each graph is the predicted confidence score of the action class where GT is the ground truth.