Two-Stream Consensus Network for Weakly-Supervised Temporal Action Localization: Supplementary Material

Yuanhao Zhai[1], Le Wang[1]*, Wei Tang[2], Qilin Zhang[3], Junsong Yuan[4], and Gang Hua[5]

[1] Xi'an Jiaotong University, Xi'an, Shaanxi, China
[2] University of Illinois at Chicago, Chicago, IL, USA
[3] HERE Technologies, Chicago, IL, USA
[4] State University of New York at Buffalo, Buffalo, NY, USA
[5] Wormpex AI Research, Bellevue, WA, USA

In this supplementary material, we first provide more implementation details of the proposed Two-Stream Consensus Network (TSCN). Then additional experimental results are provided to further analyze the efficacy of the proposed pseudo ground truth.

1 More Implementation Details

We have exploited two network backbones, *i.e.*, UntrimmedNet [4] and I3D [2], to extract features from RGB frames and optical flow. Following respective standard pipeline [4, 2], we use the original FPS of a video and a fixed FPS of 25 for UntrimmedNet and I3D, respectively. The optical flow is estimated via the TV-L1 algorithm [6].

We use a set of temporal convolutional layers to transform the original features to a set of new features. Specifically, we use 1 hidden layer for the THUMOS14 dataset [3] and 3 hidden layers for the ActivityNet datasets [1]. Each hidden layer consists of 1024 convolutional kernels with a temporal kernel size of 3 and a stride of 1. Zero padding is used to retain the dimensions of temporal location outputs. Group Normalization [5] with a group size of 32 and ReLU are used between any two consecutive hidden layers. A dropout layer with a dropout rate of 0.7 is added before the classification layer.

2 More Experiements

We conduct additional ablation studies to analyze the efficacy of the proposed pseudo ground truth. The set of ablation studies is also conducted on the testing set of the THUMOS14 dataset with UntrimmedNet features.

Ablation Study on the Fusion Parameter β. β is an important hyperparameter controlling the relative importance between the RGB stream and the flow stream at late fusion, and thus influences the quality of the pseudo ground

* Corresponding author.

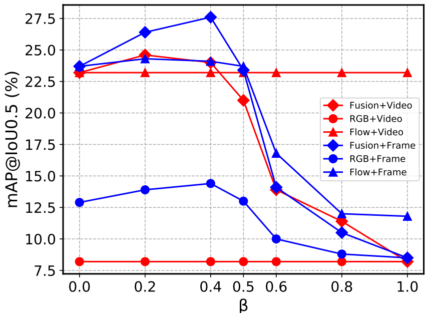

Fig. 1: Comparison of models trained with different values of the fusion parameter β on THUMOS14 testing set. "Video" denotes only video-level supervision is leveraged during training, and "Frame" denotes the frame-level pseudo ground truth is also leveraged during training

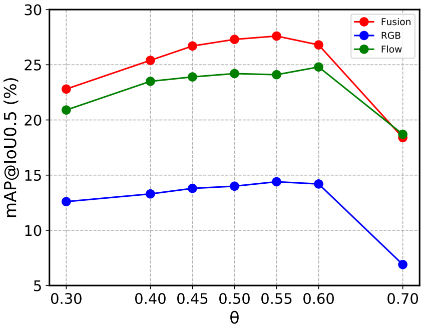

Fig. 2: Comparison of models trained with hard pseudo ground truth generated using different values of the threshold parameter θ on THUMOS14 testing set

truth. As shown in Fig 1, with only video-level supervision, the late fusion result outperforms both individual streams if the more precise flow stream outweighs (*i.e.*, $\beta < 0.5$) the less precise RGB stream. With the pseudo ground truth as frame-level supervision, the performance of both individual streams improves when β is smaller than 0.5. This further leads to better performance of their lately fused result. However, the pseudo ground truth does not always improve the two streams. As mentioned before, we set the threshold parameter θ the same as the action proposal generation threshold 0.5, which means the fusion result generated with only video-level labels is the same as the frame-level pseudo ground truth. Therefore, when β is larger than 0.5, the pseudo ground truth significantly degrades the performance of the flow stream. By contrast, the RGB stream achieves better performance under all values of β, because the pseudo ground truth consistently outperforms the RGB stream.

Ablation Study on the Threshold Parameter θ. The threshold parameter θ in the hard pseudo ground truth generation has significant impact on the quality of the pseudo ground truth. A larger θ will generate more and longer action proposals, while a smaller θ will generate fewer and shorter action proposals. We evaluate the models trained with the pseudo ground truth under different values of the threshold parameter θ, and plot the results in Fig 2. The RGB stream and flow stream achieve their best performance when θ is equal to 0.55 and 0.6, respectively. The fusion result achieves the best performance when θ is equal to 0.5 and 0.55, where both streams achieve high performance.

Ablation Study on Using T-CAM as Pseudo Labels. In the proposed method, we only use the attention sequence to generate the pseudo ground truth,

Table 1: Comparison of using T-CAM and attention sequence as pseudo ground truth on THUMOS14 testing set

Attention	T-CAM	mAP@IoU (%)		
		0.3	0.5	0.7
-	-	40.9	24.0	8.2
-	✓	41.8	24.6	7.7
✓	-	**45.0**	**27.6**	**10.2**
✓	✓	44.8	27.4	**10.2**

Table 2: Comparison of attention normalization loss in early fusion network on THUMOS14 testing set

Attention Norm	mAP@IoU (%)		
	0.3	0.5	0.7
-	28.5	14.5	4.9
✓	36.8	19.9	6.8

and the T-CAM is only used for scoring the action proposals. This is because the T-CAM is guided by the attention: we use an attention-weighted pooled feature for action classification, therefore the T-CAM corresponds with the attention. We also tried using T-CAM as pseudo ground truth, and the results are listed in Table 1. The results reveal that only using T-CAM as pseudo labels has little effects. And using both attention and T-CAM as pseudo labels has similar performance with using only attention as pseudo labels.

Ablation Study on Attention Normalization Loss in Early Fusion. We tested the attention normalization loss in early fusion framework, and the results are listed in Table 2. The performance improvement demonstrates the validity of the proposed attention normalization loss in both early and late fusion models.

Precision Recall Curve. We plot the per category Precision Recall (PR) curves obtained with and without our frame-level pseudo ground truth in Fig 3. For most of the categories, the frame-level pseudo ground truth greatly improves the precision, which further verifies the proposed pseudo ground truth helps eliminate false positive action proposals.

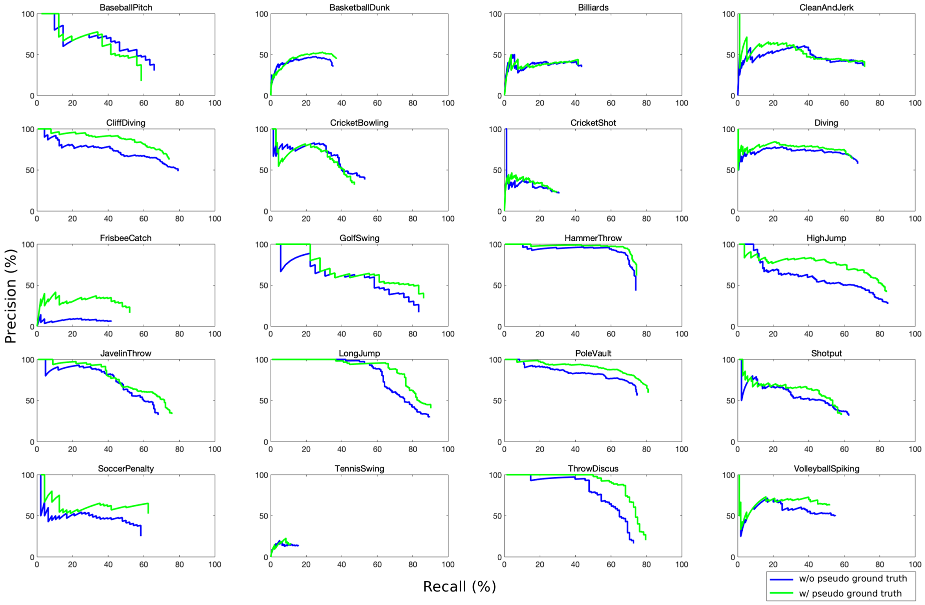

Fig. 3: Comparison between the models trained with and without the frame-level pseudo ground truth. Per category Precision Recall (PR) curves of the late fusion results on the THUMOS14 testing set are plotted. The precision and recall are calculated using an IoU threshold of 0.3. The horizontal and vertical axes correspond to recall and precision, respectively. The area enclosed by the PR curve and both axes is the Average Precision (AP)

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
3. Jiang, Y.G., Liu, J., Zamir, A.R., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: Thumos challenge: Action recognition with a large number of classes (2014)
4. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017)
5. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision. pp. 3–19 (2018)
6. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: Proceedings of the 29th DAGM Conference on Pattern Recognition. p. 214–223 (2007)