

Supplementary Materials: Revisiting Unsupervised Temporal Action Localization: The Primacy of High-Quality Actionness and Pseudolabels

Anonymous Authors

1 MORE METHOD DETAILS

1.1 Training Details

1.1.1 Overall Training Process. Compared to the conventional "clustering and training" iterative process of UTAL methods, the training pipeline of COPL simply includes the following three modifications: 1) replacing the class-agnostic attention with our HAS attention for global video feature aggregation, 2) replacing the entire pseudolabeled video set P_o with our refined pseudolabeled video set P and unlabeled video set U , and 3) adopting our IOCNet, which enhances the action awareness with multiple consistency constraints, for model training. The whole training process is presented in Algorithm 1.

1.1.2 Data Augmentation Details. Our IOCNet follows a Teacher-Student structure, wherein both the teacher and student networks share an identical architecture. The student network is trained using the refined pseudolabeled set P and the unlabeled set U , with the weights ϕ of the teacher model updated via Exponential Moving Average (EMA) [7] from the corresponding weights ϕ' of the student.

During the training, the RGB and optical flow features X^R and X^F are fed into the IOCNet, while the teacher network receives the original features and the student network receives the augmented features. Specifically, we employ the temporal feature shift technique [9] as our data augmentation method in IOCNet. This technique involves selecting a certain proportion of channels and then shifting half of them to the left and the other half to the right. This channel mixing concept has also been utilized in other approaches such as [2, 10].

Note that our focus is not on designing novel data augmentation techniques. Instead, we have chosen this specific augmentation approach to better leverage the information from the unlabeled video set U within the teacher-student framework, thereby improving the IOCNet's awareness of temporal actions.

1.2 Inference Details

In the inference stage, following our PLRI module, we obtain the refined pseudolabeled video set P that is grouped into C clusters. To facilitate the evaluation of the unsupervised representation learning ability of our COPL framework, we map each cluster to groundtruth labels [1, 3, 11], allowing comparison of the mean Average Precision (mAP) metric with prior work including weakly supervised methods.

Since we've established the correspondence between pseudolabels and groundtruth labels, we can adopt an inference process akin to the WTAL method. Specifically, for a given test video, we compute h_{cs} as outlined in Eq. (14) and subsequently select the top- T instances with the highest scores. We then utilize CAS to denote the probability of each position corresponding to the occurrence of

Table A: Clustering performance comparison with SOTA methods on THUMOS'14

| Method | Purity | NMI | ARI |
|----------|--------------|--------------|--------------|
| TCAM [1] | 0.780 | 0.811 | 0.612 |
| APSL [3] | - | 0.821 | 0.639 |
| COPL | 0.870 | 0.867 | 0.742 |

actions in different categories. Following recent works [4, 6], each selected instance's position in CAS represents the likelihood of the snippet containing the corresponding action occurrence. We then employ a series of thresholds to filter out snippets with probabilities consistently greater than the threshold θ , forming consecutive sequences as candidate proposals. Finally, we apply class-wise Non-Maximum Suppression (NMS) [5] to eliminate proposals with high overlap.

2 ADDITIONAL EXPERIMENTAL RESULTS

2.1 Clustering Performance Comparison

Table A presents the clustering results for all samples in THUMOS'14 using different methods during the final cycle, with purity, Normalized Mutual Information Score (NMI), and Adjusted Rand Index (ARI) serving as evaluation metrics. It is worth noting that the UGCT method does not report its clustering performance and has not made its code publicly available. Our reimplementation falls significantly short of the published results for UGCT in terms of localization performance; hence, we do not include a comparison of its clustering results here.

As depicted in the table, our COPL consistently outperforms TCAM and APSL across all metrics. The findings in Table 3 of our paper have already validated that incorporating the HAS module substantially enhances global video representations, thereby improving clustering performance. The results in Table A further underscore the pivotal role of PLRI and IOCNet in fortifying our HAS module, leading to enhanced clustering performance.

2.2 Qualitative Analysis of PLRI Strategy

Fig. A illustrates the distributions of confidence scores s_i for all videos from the THUMOS'14 and ActivityNet v1.2 datasets across iterations. As in Fig. A, varying confidence scores can be observed among different videos. Consequently, upon setting a threshold δ , our PLRI strategy proves effective in filtering out pseudo-labeled samples with low confidence, indicative of lower intra-cluster cohesion and inter-cluster separation, during the training process. Furthermore, as the iterations progress, a noticeable upward trend

Algorithm 1 Pipeline of our COPL framework

Input: Unlabeled video set V ; Total action categories C ;
Output: IOCNet model $\phi(\theta; \cdot)$ with parameters θ ;

- 1: **Initialize:** the IOCNet $\phi(\theta; \cdot)$; hyperparameters γ, α, δ .
- 2: **for** X_i in V **do**
- 3: Extract X^R and X^F for X_i with pre-trained extractors.
- 4: Feature concatenation: $X = [X^R, X^F]$
- 5: **end**
- 6: **for** n in $[1, iteration_num]$ **do**
- 7: /* The HAS module generates the global feature F . */
- 8: **for** X_i in V **do**
- 9: Obtain A^{HAS} with A_{cs} and A_{ca} based on Eq. (1) to Eq. (4)
- 10: Obtain video global features for X_i :

$$F = L_2 Norm(X^T A^{HAS})$$
- 11: **end**
- 12: /* The PLRI module for the refined pseudolabeled set P and unlabeled set U . */
- 13: Generate P_o with spectral clustering:

$$P_o = \{(X_i, y_i) | X_i \in V\} \leftarrow SpectralClustering(F)$$
- 14: **for** X_i in V **do**
- 15: Compute intra-cluster cohesion a_i and inter-cluster separation b_i with Eq. (6) and Eq. (7);
- 16: Compute confidence score: $s_i = \frac{a_i - b_i}{\max(a_i, b_i)}$
- 17: **end**
- 18: Split P_o into U and P based on Eq. (9) and Eq. (10);
- 19: /* Training of IOCNet on P and U . */
- 20: **for** m in $[1, epoch_num]$ **do**
- 21: Sample a mini-batch from U and P
- 22: Compute A_{cs} and A_{ca} for X_i from IOCNet:

$$A_{cs}, A_{ca} \leftarrow \phi(\theta; (X^R, X^F))$$
- 23: Compute unsupervised loss on U :

$$L_U = L_{CI} + \lambda_1 L_C + \lambda_2 L_{CO}$$
- 24: Compute all loss on P :

$$L_P = L_{cls} + L_{asl} + L_{CI} + \lambda_1 L_C + \lambda_2 L_{CO}$$
- 25: Update student network parameters θ^* :

$$\theta_t^* = \theta_{t-1}^* - \eta \frac{\partial(L_U + L_P)}{\partial \theta_{t-1}^*}$$
- 26: Update teacher network parameters θ via EMA:

$$\theta_t = \gamma \theta_{t-1} + (1 - \gamma) \theta_t^*$$
- 27: Update attention A^{HAS} generated by HAS module:

$$A^{HAS} \leftarrow HAS(A_{cs}, A_{ca})$$
- 28: **end**
- 29: **end**
- 30: **return** $\phi(\theta; \cdot)$

is observed in the overall confidence distribution. This trend signifies the effective improvement of intra-cluster cohesion and inter-cluster separation among all videos. This improvement is attributed

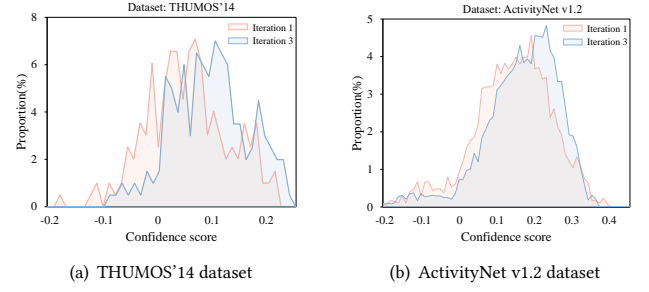


Figure A: The statistical distribution of confidence scores for P_o at iteration 0 and iteration 3, with binning at intervals of 0.01. Best viewed in color.

Table B: different δ in PLRI strategy.

| Method | δ | THUMOS'14 | | ActivityNet v1.2 | |
|----------|----------|-------------|-------------|------------------|-------------|
| | | @0.5 | @Avg | @0.75 | @Avg |
| w/o PLRI | - | 32.2 | 40.1 | 28.6 | 28.9 |
| COPL | -0.10 | 32.3 | 40.6 | 28.5 | 29.2 |
| | -0.05 | 32.6 | 41.2 | 28.7 | 29.5 |
| | 0 | 33.9 | 41.7 | 29.1 | 29.9 |
| | 0.05 | 33.3 | 41.0 | 28.9 | 29.7 |
| | 0.10 | 29.9 | 38.2 | 28.3 | 29.0 |

to the synergistic effect of our proposed modules, and it also indicates that, over the iterations, the confidence levels of video samples generally increase.

Additionally, we conducted experiments across a series of δ values in Table B. Combining those results with the confidence distributions in Fig. A, we found that selecting an appropriate δ value to filter out the proportion of pseudolabeled videos is crucial. When the filtering proportion is too low, the refined pseudolabeled set still contains some noisy samples, which adversely affects the localization performance of the COPL framework. Conversely, when the filtering proportion is too high, the available pseudolabeled video samples for training become too few, the reduction in sample quantity diminishes the localization performance.

2.3 Qualitative Analysis of HAS module

As illustrated in Fig. B, we employ t-SNE [8] to visualize video global features of six action categories within the THUMOS'14 dataset. In the figure, distinct colors represent different groundtruth action categories. A comparison between the t-SNE results with unfiltered attention in Fig. 2(a) and our proposed HAS module in Fig. 2(b) reveals that our HAS module effectively generates improved attention for video global features. For the same video snippet features, the HAS attention enhances the intra-cluster cohesion and inter-cluster separation of video global features in the feature space, consequently improving the precision of generated video-level pseudolabels.

REFERENCES

- [1] Guoqiang Gong, Xinghan Wang, Yadong Mu, and Qi Tian. 2020. Learning temporal co-attention models for unsupervised video action localization. In *Proceedings*

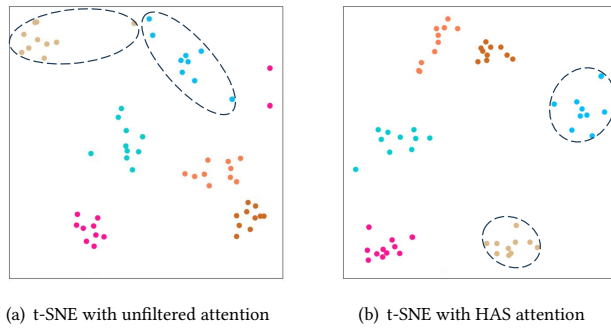


Figure B: Qualitative results of 2-dimensional t-SNE distributions of the feature space.

- [2] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. 2023. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 9819–9828.
- [3] Yuanyuan Liu, Ning Zhou, Fayong Zhang, Wenbin Wang, Yu Wang, Kejun Liu, and Ziyuan Liu. 2023. APSL: Action-positive separation learning for unsupervised temporal action localization. *Information Sciences* 630 (2023), 206–221.
- [4] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 2019. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8679–8687.
- [5] Phuc Xuan Nguyen, Deva Ramanan, and Charles C Fowlkes. 2019. Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5502–5511.
- [6] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. 2020. Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1009–1019.
- [7] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 30 (2017).
- [8] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).
- [9] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. 2021. Self-Supervised Learning for Semi-Supervised Temporal Action Proposal. In *CVPR*.
- [10] Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Xian-Sheng Hua, and Lei Zhang. 2022. Spatiotemporal self-attention modeling with temporal patch shift for action recognition. In *European Conference on Computer Vision*. Springer, 627–644.
- [11] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. 2022. Uncertainty Guided Collaborative Training for Weakly Supervised and Unsupervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).