

Supplementary Materials: TS-ILM:Class Incremental Learning for Online Action Detection

Anonymous Authors

A SUPPLEMENTARY DETAILS

Dataset details Following convention [7, 11], we evaluate our model on two publicly available datasets, which are the standard OAD datasets: THUMOS'14 [4] and TVSeries [3]. The THUMOS'14 dataset contains a large number of human daily living actions and sports actions, and its videos are sourced from the YouTube website. This dataset is divided into training, testing, and validation sets. We use its validation set for training and its testing set for evaluation. The validation set includes 200 untrimmed videos and the testing set includes 213 untrimmed videos, both of which cover the same 20 action categories, including actions such as diving, golf swing, volleyball spiking, etc. The TVSeries dataset consists of 27 untrimmed long videos containing 6 popular TV series, each about 16 hours long and around 150 minutes per series. This dataset is divided into training, validation, and testing sets, each containing at least one episode from a TV series. This dataset includes 30 action categories, such as opening door, reading, eating, etc., with the rest corresponding to background classes. In keeping with convention [2, 11], we choose 20 videos for training and evaluate on the remaining 7. For THUMOS'14, we have two settings, 10 tasks and 20 tasks, where for the former, the action categories are divided into 10 tasks, with each task containing two, and for the latter, the action categories are divided into 20 tasks, with each task containing one. For TVSeries, we have two settings, 10 tasks and 30 tasks, where for the former, the action categories are divided into 10 tasks, with each task containing three, and for the latter, the action categories are divided into 30 tasks, with each task containing one.

Training details LSTR [12] is used as our backbone, and we follow LSTR's data preprocessing process. To learn model weights, we employ an Adam [5] optimizer with weight decay of 5×10^{-5} , where the learning rate increases linearly from 0 to 5×10^{-5} during the first 2/5 of iterations and then decreases to 0 following a cosine function. Each incremental step in the training phase lasts for 25 epochs, with a batch size of 16. For a fair comparison, the training setup and backbone are the same for all methods. The loss weights γ and η are set to 0.95 and 0.45, respectively. Moreover, within the Temporal-Sensitive Exemplar Selector (TES), frames are preliminarily filtered using the herding strategy [9], followed by a further selection with the Minimum Distance Selection and storage. The ratio of the number of frames retained through these two filtering mechanisms is 2:1, meaning that frames sized at twice the replay memory are initially filtered out, followed by carefully selecting the specified number of frames according to the replay memory to save. In order to compare our method with existing continual learning approaches, we have re-implemented each algorithm. For regularization methods, MAS [1] and EWC [6], the hyperparameter λ_{reg} yields the best results when set at 5×10^2 and 3×10^4 , respectively.

Table 1: Sensitivity of the performance of TS-ILM to γ and η on THUMOS'14 with 10 steps. Default settings are marked in gray .

γ	η	THUMOS'14		TVSeries	
		mAP \uparrow	Forget \downarrow	cAP \uparrow	Forget \downarrow
0.95	0.4	53.81%	33.76%	76.74%	11.90%
0.95	0.5	53.66%	33.51%	77.10%	11.31%
0.95	0.45	54.03%	33.13%	77.26%	11.71%
1	0.45	53.01%	34.10%	76.62%	12.20%
0.9	0.45	52.71%	34.73%	76.88%	12.56%

Table 2: Sensitivity of the performance of TS-ILM to α and β on THUMOS'14 with 10 steps. Default settings are marked in gray .

α	β	THUMOS'14		TVSeries	
		mAP \uparrow	Forget \downarrow	cAP \uparrow	Forget \downarrow
0.1	0.5	54.03%	33.13%	77.26%	11.71%
0.5	0.1	50.91%	40.40%	75.10%	17.54%
1	0.05	49.75%	45.59%	74.98%	17.13%
0.05	1	45.13%	63.03%	75.90%	15.67%
0.2	0.25	53.10%	32.67%	77.04%	11.93%
0.25	0.2	52.74%	34.09%	76.90%	12.50%

B ADDITIONAL RESULTS

In Figure 1, we demonstrate the performance of various models at each incremental step on additional datasets. TS-ILM achieved higher accuracy in most incremental steps, indicating its strong capability to preserve past knowledge. At the same time, we can observe that on both splits of the TVSeries dataset, the decline rate in model performance is much lower than on the THUMOS'14 dataset. We conjecture this is because the latter possesses more complex action patterns, leading to greater potential confusion among actions.

C ADDITIONAL ABLATIONS

Effect of Balance Weights We further discussed the sensitivity of the balance weights γ and η for each term in the final loss L on THUMOS'14 with 10 steps and TVSeries with 10 steps. Table 1 shows the performance of TS-ILM under various combinations of γ and η . We found that the performance under the combination of $\{\gamma = 0.95, \eta = 0.45\}$ is consistently superior to other combinations. Although the combination of $\{\gamma = 0.95, \eta = 0.50\}$ resulted

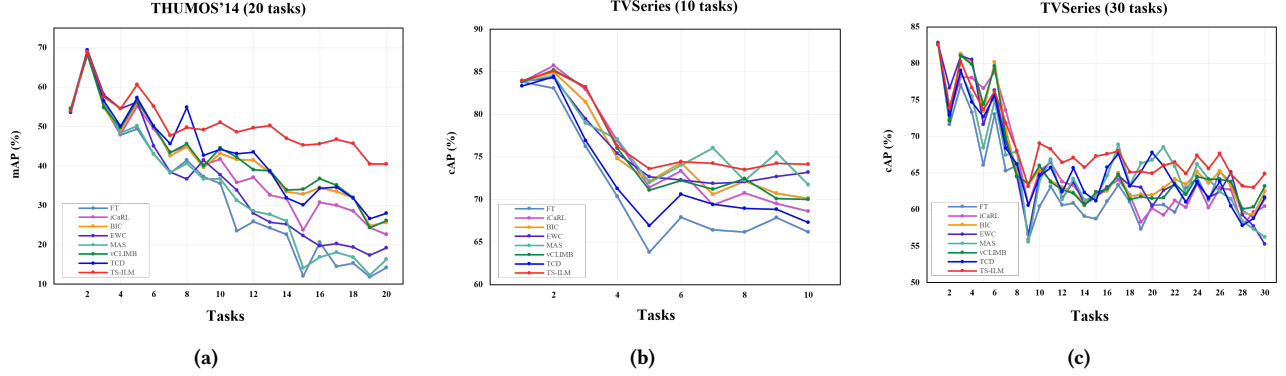


Figure 1: The performance of various methods on more datasets at each incremental step: (a) THUMOS'14 datasets with 20 steps; (b) TVSeries datasets with 10 steps; (c) TVSeries datasets with 30 steps. In most incremental steps of these three datasets, TS-ILM achieved higher accuracy, indicating its strong ability to retain past knowledge.

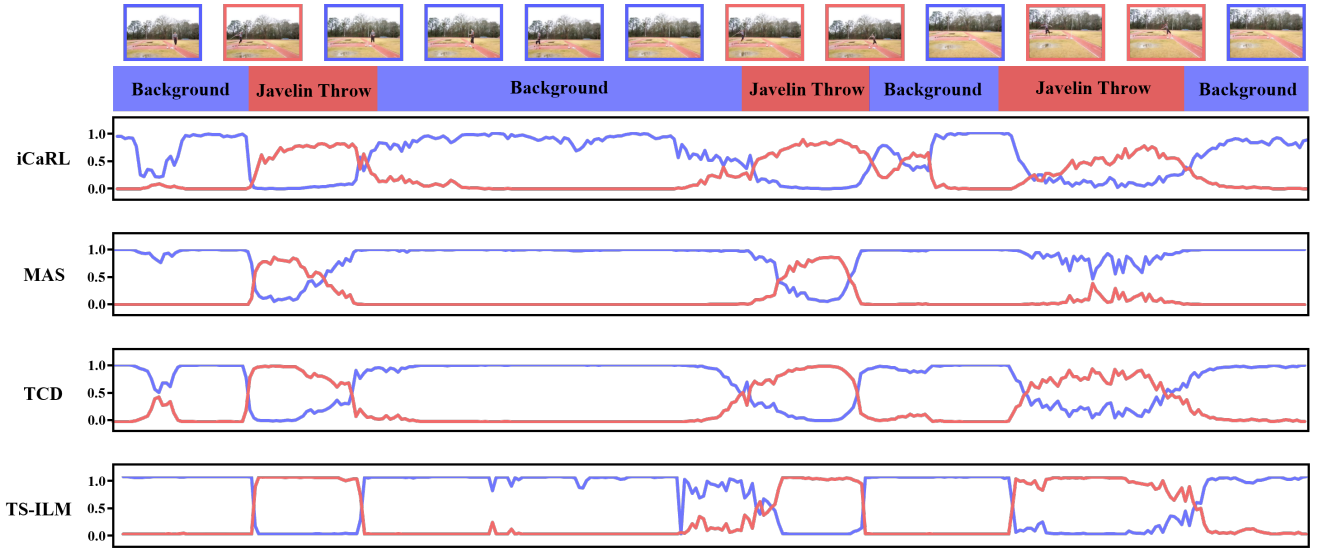


Figure 2: Qualitative analysis of the proposed TS-ILM and the CIL methods applied in the image and video domains. The bars in different colors represent the true categories, while the lines represent the action scores of the method.

in a slightly lower forgetting rate on the TVSeries with 10 steps, other metrics were far less satisfactory than those of our ultimately chosen combination.

Effect of selection ratio in TES We further discussed the sensitivity of the selection proportions α and β of the two modules in TES under the situation where the replay memory is 5% of the total data volume on THUMOS'14 with 10 steps and TVSeries with 10 steps. Table 2 shows the performance of TS-ILM under various combinations. We found that the performance under the combination of $\{\alpha = 0.1, \beta = 0.5\}$ is generally better than other combinations. We speculate that this is because it is necessary to first select enough frames through the herding strategy [9] and then screen out more time-related frames with the Minimum Distance Selection. If the proportion of the former is too large, it may lead to the resulting frames not being representative. If the proportion of the former is

too small, it may lead to a weakening of the temporal correlation of the finally saved frames.

D ADDITIONAL QUALITATIVE ANALYSES

Figure 2 visualizes the action scores obtained by different methods alongside the corresponding video segments. These action scores were inferred from videos in the dataset of the 10-th task after training on 10 tasks on THUMOS'14. The “Javelin Throw” category showcased in the figure belongs to the training set of the 7-th task and has not been included in later training. The results show that our proposed TS-ILM has stronger retention of previous action categories, significantly mitigating the issue of catastrophic forgetting, compared to the regularization method MAS [1] used in the image domain, the exemplar replay method iCaRL [9] also in the image domain, and the class incremental learning method TCD

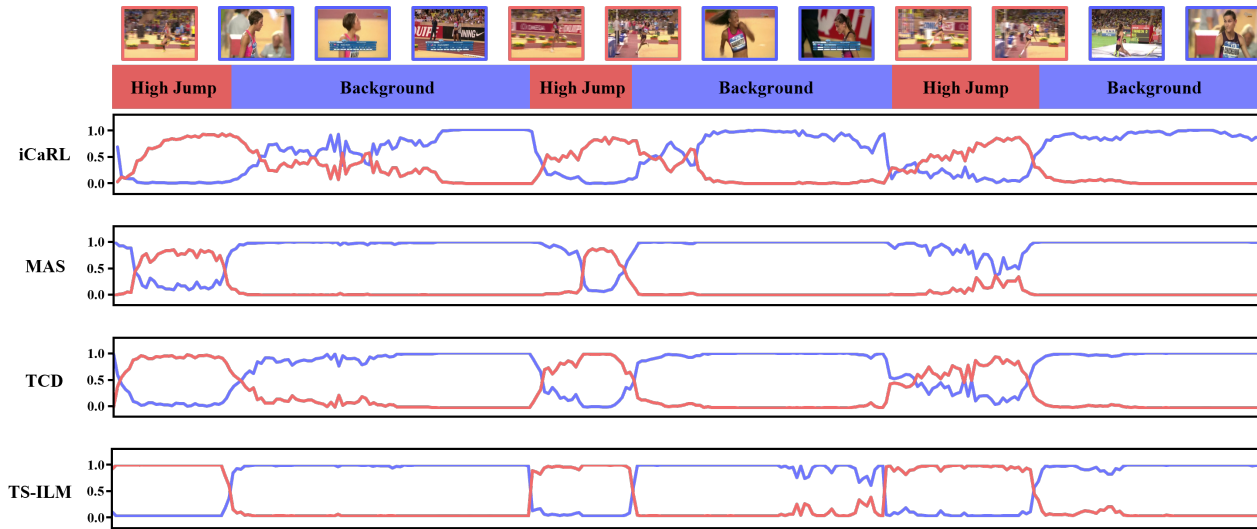


Figure 3: Qualitative analysis of various methods on video clips that include actions from other categories. The bars in different colors represent the true categories, while the lines indicate the action scores of the method.

[8] implemented in the video domain. Figure 3 visualizes video clips containing actions from various categories as well as their corresponding action scores. The “High Jump” displayed in the figure belongs to the training set of the 6-th task, and it has not reoccurred in later training phases. It can be observed that our TS-ILM exhibits robust retention capabilities for previous action categories, regardless of the type of action.

E FAILURE ANALYSES AND FUTURE WORK

Despite our proposed TS-ILM having effectively overcome the problem of catastrophic forgetting compared to the baseline, its performance still needs improvement in some cases. Specifically, TS-ILM sometimes confuses different actions, resulting in low confidence for the action categories depicted in the Figure 2. This could be due to an insufficient capacity to retain memories of historical actions. This theoretical and qualitative analysis leads our future work, which includes, but is not limited to: (1) Extracting more effective information from videos instead of merely saving video frames, thus expanding the amount of information preserved. (2) Designing a module to separate actions from the background, effectively distinguishing between actions with similar backgrounds. (3) Mining more knowledge from the various types of data saved.

REFERENCES

- [1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*. 139–154.
- [2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [3] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. 2016. Online action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*. Springer, 269–284.
- [4] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding* 155 (2017), 1–23.
- [5] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [6] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.
- [7] Siyu Liu, Jian Cheng, Ziyang Xia, Zhilong Xi, Qin Hou, and Zhicheng Dong. 2023. HCM: Online Action Detection With Hard Video Clip Mining. *IEEE Transactions on Multimedia* (2023).
- [8] Jaeyoo Park, Minsoo Kang, and Bohyung Han. 2021. Class-incremental learning for action recognition in videos. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13698–13707.
- [9] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2001–2010.
- [10] Andrés Villa, Kumail Alhamoud, Victor Escorcia, Fabian Caba, Juan León Alcázar, and Bernard Ghanem. 2022. vclimb: A novel video class incremental learning benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19035–19044.
- [11] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. 2021. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7565–7575.
- [12] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. 2021. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems* 34 (2021), 1086–1099.