Time-Class Cross-Attention Classifier for Exemplar-Free Continual Learning in Video Action Recognition

Anonymous submission

Abstract

In the context of video domains, continual learning has traditionally relied on data storage to prevent forgetting the knowledge learned in the previous tasks. However, due to the substantial size of data compared to images, it costs significant storage complexity and time complexity to store and select important frames. To this end, we explore methods to maintain prior information without storing or reusing data, proposing a Time-Class Cross-Attention Classifier for continual learning in video action recognition. We employ learnable class queries to compress class knowledge, and a cross-attention classifier architecture to capture the relationship between class queries and temporal information in videos. Then we transfer information from the previous cross-attention classifier when learning new tasks to preserve the necessary temporal cues for the classification of previous classes. Experimental results show that the proposed model significantly improves performance in scenarios regardless of whether data reuse is feasible or not, offering a novel perspective on continual learning in the field of action recognition. Our code will be made available.

Introduction

Video action recognition, a fundamental task in computer vision, has witnessed remarkable progress with the advent of deep learning. Deep models have demonstrated the capability to recognize a wide range of actions in videos, from simple gestures to complex human movements. However, the practical application of these models in real-world scenarios often demands the ability to learn continually, adapting to new action classes and data streams without compromising knowledge acquired from previous tasks.

Continual learning in the context of video action recognition addresses this problem, but still presents a formidable challenge. Traditional approaches have often relied on memory rehearsal, storing video frames in a memory bank and training them along with new data to maintain performance on previously learned tasks while adapting to new ones. While these methods have proven effective, they introduce a notable contrast to the core premise of continual learning, the ability to maintain prior task performance without overly relying on retraining substantial portions of the previous dataset. Memory rehearsal methods inherently involve retraining a considerable amount of the existing data, presenting a potential bottleneck in the continual learning process.

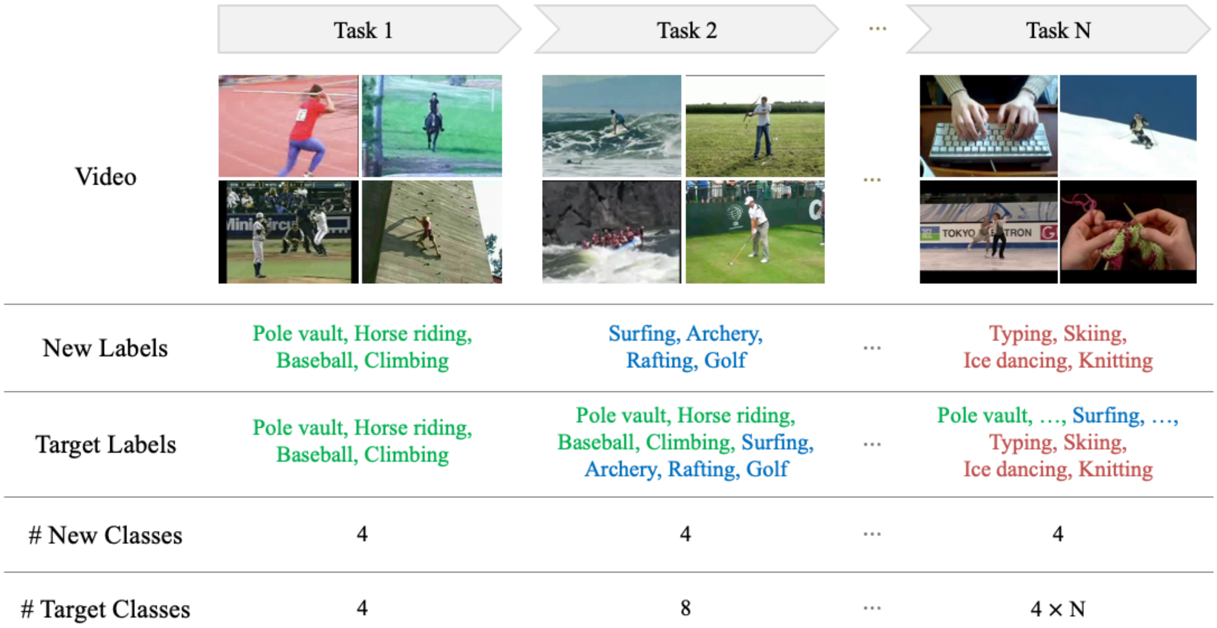
Furthermore, the practical applicability of memory rehearsal is limited in real-time systems that demand prompt responses. Such systems cannot tolerate delays arise during the selection of important frames from incoming video streams, a common issue in existing memory rehearsal-based methods. Additionally, accessing stored data is unfeasible in environments with strict privacy regulations or limited storage capacity. These constraints emphasize the need for alternative approaches that do not rely on the rehearsal of past data.

To tackle this issue, we address the exemplar-free continual learning approach in video action recognition by introducing Time-Class Cross-Attention classifier. To maintain the information of previous classes without storing data, we train class-specific queries. We then perform cross-attention between video segment features and class queries, allowing the model to capture which temporal segment of the video is significant for predicting right action class. The key innovation of our approach lies in its ability to compress class information into class queries by attending to crucial temporal cues for the class. The training of class queries is achieved by adding a lightweight cross-attention layer. As a result, any model with a video feature extractor can easily adopt our method by simply adding the learnable class queries and a cross-attention classifier, avoiding any intensive computational demands. The learned knowledge is transferred to the next task, persisting the acquired class knowledge for subsequent tasks.

Our comprehensive experiments show notable improvements in performance in scenarios where the reuse of data is not available. We demonstrate that our single-layered cross-attention classifier effectively preserves previous knowledge in the form of class queries, without relying on a Large Language Model(LLM) or pre-trained models. Our work paves a new path for continual video action recognition systems, introducing a more practical approach in the field of video continual learning.

Our contributions are as follows:

- We propose a very simple plug-and-playable module for class-incremental video action recognition without the need for storing past data. We suggest and validate a novel way to predict class logit.



	Task 1	Task 2	···	Task N
Video			···	
New Labels	Pole vault, Horse riding, Baseball, Climbing	Surfing, Archery, Rafting, Golf	···	Typing, Skiing, Ice dancing, Knitting
Target Labels	Pole vault, Horse riding, Baseball, Climbing	Pole vault, Horse riding, Baseball, Climbing, Surfing, Archery, Rafting, Golf	···	Pole vault, …, Surfing, …, Typing, Skiing, Ice dancing, Knitting
# New Classes	4	4	···	4
# Target Classes	4	8	···	4 × N

Figure 1: Description of class-incremental learning for action recognition task. In each task, a set of new action classes is introduced, and the model is required to classify all the classes seen so far. Unlike task-incremental learning, there is no provision of task information in the input.

- We demonstrate that efficient data replacement can be achieved by compressing information about video action labels into class queries.
- We introduce a new perspective on exemplar-free approaches to video continual learning and provide a baseline that serves as the starting point for this approach.

Related Works

Continual learning has seen substantial growth and development, especially in the domain of computer vision. However, video action recognition has received comparatively less attention in the continual learning research community. While action recognition has made impressive strides in general, the adaptation of continual learning techniques for this specific domain remains constrained and uncharted. In this section, we review notable works in the realm of continual learning and its application to video action recognition.

Continual Learning

Continual learning aims to enable models to learn new classes while retaining knowledge of previously learned ones. The strategies can be broadly categorized into four techniques, each aimed at mitigating the catastrophic forgetting problem, a fundamental challenge in continual learning.

Regularization-based Methods. Regularization methods focus on preventing the change of important weights learned in the previous task. EWC(Kirkpatrick et al. 2016) employs the Fisher Information Matrix to quantify the importance of weights, while LwF(Li and Hoiem 2017) distills the output of the old model when training the new task to preserve existing knowledge. This approach has been widely used in continual learning, but is often computationally heavy and exhibit limited generalization ability.(Van de Ven and Tolias 2019).

Replay-based Methods. Methods such as iCaRL(Rebuffi et al. 2016) and Gradient Episodic Memory (GEM)(Lopez-Paz and Ranzato 2017) concentrate on replaying previous data to uphold knowledge of previously learned classes. The replay-based strategy stands out as one of the most effective methods for preserving past knowledge by retaining the original training data. However, its practicality diminishes when faced with limited memory or substantial size of each exemplar.

Architectural Approach. Architectural approaches enlarge network's capacity dynamically through growing the number of network parameters in accordance with the increased amount of classes(Yoon et al. 2017; Teja and Panda 2020). DEN(Yoon et al. 2017) selectively updates parameters in the existing network and dynamically expands the trained network by assessing whether additional capacity is needed during training. CaCL(Teja and Panda 2020) compacts the expanded network efficiently by employing low-rank approximation and task-specific residual learning.

Prompting Methods. Prompting based methods(Radford et al. 2021; Thengane et al. 2022) originated with the advent of prompt learning utilizing large language model. L2P(Wang et al. 2022b) is the first attempt to adopt prompt learning using a pretrained transformer encoder in the continual learning of image classification domain. DualPrompt(Wang et al. 2022a) trains two sets of disjoint prompt pools to decouple higher level prompt spaces. ZSCL (Zheng et al. 2023) preserves Zero-Shot transfer ability in Continual Learning by leveraging the original CLIP model as a teacher in feature space.

These approaches have shown promise in alleviating catastrophic forgetting in the image domain, but may not fully address the unique demands of video action recognition. Transitioning to video domain may render them impractical in real-time systems due to the substantial storage demands associated with video data and a large number of model weights.

Video Action Recognition

In the domain of video action recognition, numerous models and datasets have been established to advance the state of the art. Works such as Two-Stream CNNs(Simonyan and Zisserman 2014), I3D(Carreira and Zisserman 2017), and TSN(Wang et al. 2016) have significantly enhanced the accuracy of action recognition. However, these models are often optimized for static datasets and single-task scenarios, making the transition to more complex, dynamic environments like continual learning and real-time applications a non-trivial challenge.

Attention Mechanisms

Concurrently, attention mechanisms have emerged as powerful tools in a diverse array of natural language processing and computer vision tasks, characterized by their capacity to spotlight interrelated information within the data. Transformer(Vaswani et al. 2017) played a pivotal role in highlighting the significance of attention mechanism. With the advent of the Vision Transformer(ViT)(Dosovitskiy et al. 2020), Transformer has been applied in various ways in the vision domain. DETR(Carion et al. 2020) introduced learnable query embeddings that attend to each object for object detection in images, and Mask2Former(Cheng et al. 2022) proposed masked attention to support universal image segmentation. In the context of action recognition, divided attention has been employed separating spatial attention and temporal attention.

Continual Learning for Action Recognition

A limited number of studies have explored the challenges of continual learning in video action recognition. TCD(Park, Kang, and Han 2021) can be credited as one of the pioneering attempts at introducing continual learning into the realm of action recognition. This method proposed channel-wise importance to select feature maps that have a greater impact on previous tasks. vCLIMB(Villa et al. 2022b) introduced the first benchmark for video continual learning, addressing the challenge of storing videos in memory through the incorporation of a temporal consistency loss. It further emphasized the prominence of rehearsal-based methods as top performers, where a subset of video frames is retained in the memory bank. However, such approaches suffer from the need to select and store frames in the replay memory, incurring substantial time and storage costs.

SMILE(Alssum et al. 2023) aimed to alleviate this burden by preserving one frame per video and subsequently generating a video with repetitive copies of this image for training. PIVOT(Villa et al. 2022a) utilized a pre-trained image and text encoder derived from CLIP to harness knowledge from a large-scale embedding space to enable zero shot transfer of knowledge. However, the training procedure becomes complex as prompt-based methods require the training of a task identifier to lookup task-specific prompts (Villa et al. 2022a; Wang et al. 2022b). Even though they aims to leverage knowledge on the unknown set using a large pre-trained model, they still rely on the replay of exemplars and presents large performance drop without directly storing class data. Hence, we aim to present an approach with minimal time and storage requirements.

Although there were some noteworthy contributions, the limited body of research in the intersection of continual learning and action recognition underscores the need for novel approaches and solutions. Our proposed Time-Class Cross-Attention classifier aims to address this gap by integrating elements of continual learning with video specific attention mechanisms, presenting an innovative perspective on exemplar-free continual learning in the context of action recognition.

Method

We present the Time-Class Cross-Attention classifier and its distillation protocol. Our method is designed to adapt to new action classes while retaining knowledge of previously learned ones, all without the need for extensive data storage or memory rehearsal.

Problem Statement

We focus on class-incremental learning, which aims to train a single model with new class set that arrives while maintaining the performance on previous class sets. A single model Θ is trained through a task sequence $T = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_t, \ldots\}$, each of which are trained in different time step t. Each task $\mathcal{T}_t$ consists of videos of distinct action classes and corresponding action label set pair, $\{\mathcal{X}_t, \mathcal{Y}_t\}$. Each pair contains the same number of new action labels. We evaluate the model's performance using test data of all the classes seen so far without any task information.

Temporal Feature Extractor

We employ TSN(Wang et al. 2016) as a base feature extractor to capture important spatio-temporal features which are crucial for recognizing action. An input video is divided into K segments, and each segment is fed into the ResNet(He et al. 2016) module. Rather than passing through a consensus module that original TSN has, we train the cross attention classifier to find weights for each temporal segment. We select ResNet34 based TSN for a fair comparison with previous research.

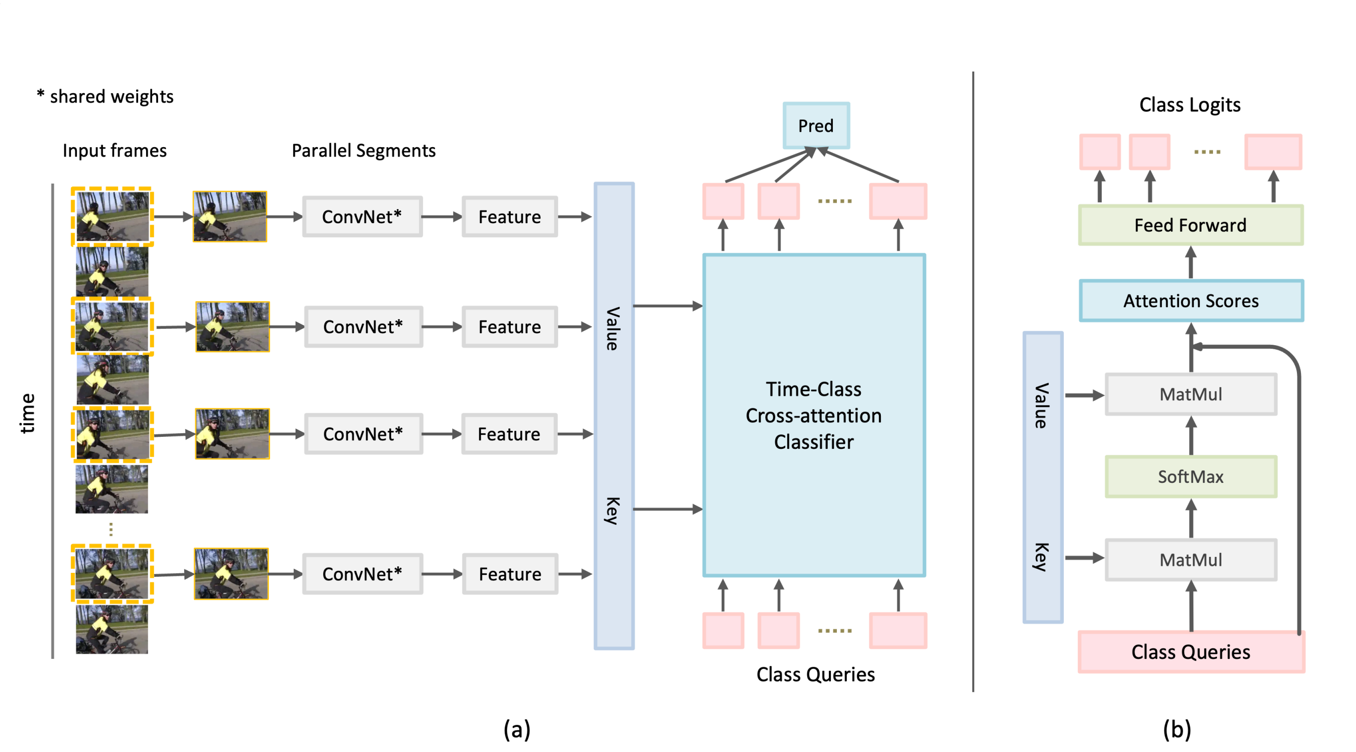


Figure 2: (a) Overall architecture of our approach. Learnable class queries are randomly initialized, and key and values are transformed from temporal features. (b) Details of Time-Class Cross Attention classifier. The cross attention module computes attention between class queries and temporal features, to produce output logits.

Time-Class Cross Attention (T-CCA) Classifier

To achieve continual learning without data storage, we design a classifier tailored for training learnable class queries. Cross attention is adopted as the core architecture of the classifier, to highlight the important temporal segment for predicting the right class. The module calculates attention between class queries and temporal features, followed by a feed forward layer which outputs a class logit for each query. The logits are passed through the softmax function to generate the final class prediction.

Let's denote the inputs for the current task t as x_t. x_t is divided into K segments, and they are passed through the current feature extractor F_t to extract K features, denoted as f_t. f_t is transformed into K keys k_t and K values v_t. At the beginning of the first task, a number of learnable class queries are dynamically added and randomly initialized, corresponding to the classes appearing in the current task q_t. Then the cross attention module computes attention s_t between each key, values and learned class queries so far, to capture essential spatio-temporal cues from video sequences for classifying action.

$$s_t = Softmax(q_t \cdot k_t^T \sqrt{d_k}) v_t \tag{1}$$

$$a_t = w_t \dot{(} s_t + q_t) \tag{2}$$

$$l_t = FeedForward(a_t) \tag{3}$$

Attention score a_t is calculated by multiplying learned weights of each temporal segment and s_t. Original query feature is added here for residual connection. $a_t \in R^{B \times C_{1:t} \times D}$ is transformed to class logits $l_t \in R^{B \times C_{1:t} \times 1}$ through a feed-forward layer, where B, C, and D each stands for the batch size, number of known classes, and hidden dimension. Feed forward layer compresses the attention output into class logits. Then $\hat{y_t}$ is outputted as a final prediction considering the class with the highest logit as a correct label.

The rationale behind computing attention between class queries and temporal segments stems from the varying importance of each segment when determining action class. In contrast to the image domain, the occurrence of an action is confined to a subset of the entire video as described in Figure 3. Within this subset, specific segment of the video play a pivotal role in determining the action class. Our classifier is designed to enable queries to attend to these crucial temporal segments throughout the video. By first attending to different temporal segments in parallel, and then find weights based on their contribution to the model's performance. This procedure incorporates temporal information into class logit, enabling the classifier to discern significant temporal features crucial for predicting the action class.

To train the classifier and obtain correct predictions from the logits, temporal classification loss L_{cls} is applied. The Temporal Classification Loss minimizes the cross-entropy between the predicted class probabilities and the ground

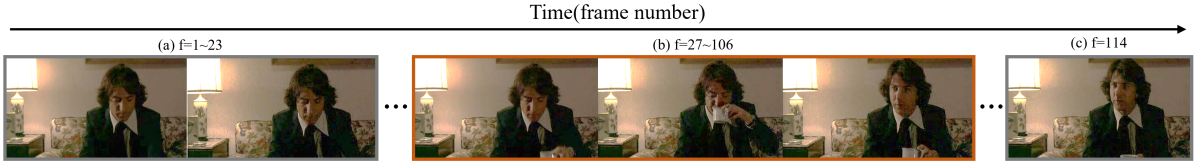


Figure 3: An example data of class 'Drink' in HMDB51. The video data starts with a man just sitting on the couch, and drinking starts from 27th frame. After the action, man stares someone without drinking or holding a cup. The key segment for predicting the provided label only corresponds to a subset of the entire input frames(Specifically, frames 27 to 106). This shows the evident example of the case why we choose to attend the class query to the segments of video divided along the temporal axis, to capture important temporal feature.

truth y_t, guiding the model to improve its action recognition capabilities. While learning proper logit distribution, class queries are trained to produce discriminative logit with each other.

$$L_{cls} = -\sum_{i=1}^{C} y_{t,i} \log \hat{y}_{t,i} \tag{4}$$

As queries are learnable, they are updated during the training process. The cross-attention classifier facilitates the simultaneous learning of important temporal features and class-specific information. The detailed structure of the module is depicted in the Figure 2.

Temporal Feature Transfer (TFT)

To ensure that knowledge of previously learned classes is retained, the weights of cross-attention module and class queries are transferred from the previous cross-attention classifier during the learning of new action classes. The Temporal Distillation Loss L_{dist} serves a crucial role in the process, retaining previous knowledge and preserving classification capabilities for prior classes. It minimizes the KL-Divergence between the distribution of logits from the previous cross-attention module and that from the current module to balance the influence of new and prior knowledge. To further mitigate forgetting, we applied weighting term to put more weights to the previous knowledge as task sequence gets longer. The loss is calculated as:

$$L_{dist_q} = P(q_{t-1}) \cdot \log P(q_{t-1}) Q(q_t) \tag{5}$$

$$L_{dist} = W_t(P(x) \cdot \log P(x) Q(x)) + L_{dist_q} \tag{6}$$

where W_t is logarithmic weight function, $P(x)$ is the softmax output of l_{t-1}, $Q(x)$ is the softmax output of l_t in the log space, $P(q_{t-1})$ and $Q(q_t)$ are previous and current distribution of class query weights. The distillation loss serves as a regularization term to prevent significant changes for predicting previous classes. Effects of each component are measured in ablation study in the supplementary material.

With balancing two loss functions, the cross-attention classifier adapts to capture new action class-specific features while retaining the ability to recognize actions learned in previous tasks. The cross-attention module captures essential temporal features crucial for predicting the correct classes, and the class query is updated to learn distinctive features for each class. The learned knowledge is retained across tasks by constraining the classifier to generate similar logit distribution for previous action classes. The final objective function L is described with balancing hyperparameter λ_1 and λ_2 as:

$$L = \lambda_1 \cdot L_{cls} + \lambda_2 \cdot L_{dist} \tag{7}$$

Experiments

In this section, we empirically evaluate the effectiveness of the Time-Class Cross-Attention Classifier in addressing the challenges of continual learning in action recognition. We present the datasets, experimental setup, and results of our method.

Datasets

For the experiments, we use a widely recognized action recognition dataset: the UCF101(Soomro, Zamir, and Shah 2012) and HMDB51(Kuehne et al. 2011). UCF101 consists of 13.3K videos spanning 101 action classes, and HMDB51 contains 6.8K videos devide into 51 action categories from public databases. Following the previous work(Park, Kang, and Han 2021), we adopt the first split among 3 official train-test splits to evaluate our approach on in-distribution training setting.

Experimental Setup

Pre-trained with in-distribution data. Initially, we conduct experiments using the class-incremental scenarios employed in TCD(Park, Kang, and Han 2021) on UCF101 and HMDB51 datasets with in-distribution pretraining. For UCF101, we first train the model with 51 classes and divide the remaining 50 classes into 5, 10, and 25 tasks for the class-incremental step. For HMDB51, the model is initially trained with 26 classes, and the remaining 25 classes are divided into 5 and 25 tasks. We employ ResNet34 based TSN for UCF101, and train the entire network for 20 epochs with a batch size of 32. ResNet50 based TSN is trained for 20 epochs with a batch size of 16 for HMDB51, to make a fair comparison with previous works.

Without pre-training in-distribution data. Following the vCLIMB benchmark(Villa et al. 2022b), we test various models on the UCF101 dataset without in-distribution pre-training. We execute experiments both with and without memory for this scenario. We evenly divide the entire set of classes into 10 and 20 tasks. To assess the performance of T-CCA, Average Accuracy(ACC) and Backward Forgetting (BWF) are measured. For experiments based on PIVOT, we train the model for 40 epochs with a batch size of 50, utilizing the SGD optimizer with a constant learning rate of 0.01. For iCaRL, we train the model for 40 epochs with a batch size of 50, utilizing the SGD optimizer with a constant learning rate of 0.01. For the memory bank, we follow the information given by each paper.

Evaluation Metrics

Following previous works, we employ the following evaluation metrics to asses the performance of our method.

Average Accuracy(ACC). Average Accuracy is a standard metric used to measure the overall accuracy of a model across all seen tasks. It is calculated as the average of accuracies achieved on individual tasks after finishing training on the current task.

$$ACC = \frac{1}{t}\sum_{i=1}^{t} acc_i \tag{8}$$

where t is the total number of tasks and acc_i is the accuracy of the trained model on the i-th task. Higher ACC means better performance.

Backward Forgetting (BWF). Backward Forgetting quantifies the extent to which a model forgets knowledge of previously learned tasks when adapting to new ones. It is computed as the difference in accuracy for the last task between models trained with and without considering previous tasks.

$$BWF = \frac{1}{t-1}\sum_{i=1}^{t-1} acc_{i,i} - acc_{t,i} \tag{9}$$

where $acc_{i,j}$ represents the accuracy of j-th task after training i-th task. Lower BWF means less forgetting, thus represents better performance.

Exemplar-free Baseline

In the absence of existing approaches for exemplar-free continual learning in action recognition tasks, we select three prior works with distinct approaches, Time Channel Distillation(TCD)(Park, Kang, and Han 2021), PIVOT(Villa et al. 2022a), and iCaRL(Rebuffi et al. 2016), and entirely remove the sections involving memory rehearsal to assess exemplar-free performance. Results are presented in Table 1 and Table 2 as TCD$_{EF}$, PIVOT$_{EF}$, and iCaRL$_{EF}$.

TCD(Park, Kang, and Han 2021) conducts pre-training on the half set of the entire classes before incrementally learning the remaining classes. To assess the efficacy of this approach in an exemplar-free scenario, we omit the use of a memory bank and class-balanced tuning, while retaining all other training details including time channel distillation. The exemplar-free performance of this baseline is presented in Table 1. TCD originally achieves an average accuracy of 74.89% with five rounds of incremental tasks, but the accuracy drops to 31.5% when we restrain using exemplar. To preserve the learned information without utilizing previous data, we integrate our methodology, the cross-attention classifier and learnable class query, achieving a performance improvement of 10.73% in the 10 classes 5 tasks setting. It consistently excels in scenarios involving 10 and 25 incremental tasks, demonstrating improvements of 8.01% and 3.87%, respectively. In the HMDB51 dataset, our method also shows improvement, indicating the ability to mitigate forgetting without a memory bank. By adopting weighting parameter to distillation loss, we could effectively enhance performances in 25 tasks setting of both dataset.

We also validate our non-exemplar approach on PIVOT(Villa et al. 2022a) and iCaRL(Rebuffi et al. 2016), following the protocol of vCLIMB(Villa et al. 2022b) benchmark. It evenly divides the entire set of classes into 5 or 10 classes per step, and experiments on 10 or 20 steps without in-distribution pre-training. We exclude the memory bank to assess the models under the non-exemplar scenario. PIVOT leverages large pretrained vison-language models(VLM), CLIP(Thengane et al. 2022), achieving a high performance of 94.8% when data is stored. However, it drops to 26.19% without storing data, showing the fact that leveraging VLM has been insufficient for retaining previous knowledge without any data storage. By incorporating class queries and the cross-attention classifier into this non-exemplar baseline, we achieve 6.45% and 3.76% improvement on 10 and 20 task experiments. iCaRL$_{EF}$ shows more promising results even without storing data, showing around 12% higher accuracy than PIVOT$_{EF}$ on 10 incremental tasks setting. Our method gains 7.89% improvement in accuracy than the non-exemplar version of iCaRL on 10 tasks, and 8.39% on 20 tasks. Thus, T-CCA has proven to be an effective approach for remembering and maintaining information on previous classes without the need for data storage.

Results with Exemplars

While we design our method to address the need for non-exemplar approaches, we expand our methodology into environments where data rehearsal is allowed to demonstrate its efficacy as a continual learning method. We plug our cross-attention classifier and class query into the original versions of TCD, PIVOT, and iCaRL for the experiments.

First, we experiment with iCaRL, a representative rehearsal-based method that stores data to maintain performance on previous classes. Our approach results in performance improvements of 5.51% and 4.07% for 10 and 20 tasks, recording 86.48% and 80.66% respectively. In the case of PIVOT, one of the state-of-the-art methods, the performance is already close to the oracle yielding 94.80% of average accuracy when data is stored. Our method shows a slight increase in average accuracy on PIVOT, and notable enhancement in backward forgetting. There is a 1.11% and 0.95% decrease in backward forgetting(BWF), which means the added module helps maintaining previous knowledge, fulfilling the intent of adopting class query. The amount can be considered significant given that the performance had al-

Table 1: Results on UCF101 and HMDB51 with in-distribution training. In each scenario, model is pre-trained on the half set of the classes, followed by incremental learning on the remaining class set.

		UCF101			HMDB51	
	Method	$10 \times 5tasks$	$5 \times 10tasks$	$2 \times 25tasks$	$5 \times 5tasks$	$1 \times 25tasks$
Exemplar-free	TCD$_{EF}$	31.53	20.12	8.41	17.60	7.62
	T-CCA$_{EF}$	42.36	28.13	12.28	22.30	10.22
With Exemplars	UCIR	74.31	70.42	63.22	44.90	37.04
	PODNet	73.26	71.58	70.28	44.32	38.76
	TCD	74.89	73.43	72.19	45.34	40.07
	T-CCA	80.34	78.86	76.63	47.82	42.06

Table 2: Results on UCF101 without in-distribution training. Following vCLIMB, we split the entire dataset into 10 and 20 tasks.

		10 Tasks		20 Tasks	
	Method	ACC↑	BWF↓	ACC↑	BWF↓
Exemplar-free	EWC	9.51	98.94	4.71	92.12
	MAS	10.89	11.11	5.90	5.31
	PIVOT$_{EF}$	26.19	62.32	14.33	56.58
	PIVOT$_{EF}$+Ours	32.64	72.74	18.09	66.23
	iCaRL$_{EF}$	36.50	68.02	28.79	73.21
	iCaRL$_{EF}$+Ours	44.39	35.12	37.18	67.52
With Exemplars	BiC	78.16	18.49	70.69	24.90
	iCaRL	80.97	18.11	76.59	21.83
	iCaRL+Ours	86.48	12.24	80.66	17.74
	PIVOT	94.80	3.89	93.70	4.77
	PIVOT+Ours	94.93	2.78	94.72	3.82

ready reached the oracle. Due to the inability to train class queries in the middle of the prompt learning module of PIVOT, we simultaneously train class queries using video features alongside prompt learning. We jointly optimize the classification loss for the cross-attention classifier with the original training loss. The result is summarized in the lower part of Table 2.

We also validate our method in an in-distribution training environment and compare the result with TCD(Park, Kang, and Han 2021). On the UCF101 dataset with a 10-classes & 5-tasks configuration, our method achieves a performance of 80.34%, surpassing the previous method by 5.45%. In a 5-classes & 10-tasks setup, we observe an accuracy gain of 5.43% reaching 78.86%, and in a 2-classes & 25-tasks setup, the accuracy increases by 4.44% reaching 76.63%. On the HMDB51 dataset, we obtain a performance improvement of 2.48% and 1.99% on 5-classes & 5-tasks and 1-classes & 25-tasks settings respectively. While there is limitation that improvement of performance in the 25-tasks environment is restricted, our method consistently demonstrates performance enhancements across all tested settings, showing its robust effectiveness. Those results and comparisons with the previous method are summarized in Table 1.

Conclusion

Our Time-Class Cross-Attention classifier presents a straightforward yet promising approach for exemplar-free continual learning in action recognition. Through the incorporation of class queries and a cross-attention layer, we demonstrate a way to store class information without storing extensive video sequences or selecting frames. We have taken the initial step in video continual learning to apply non-exemplar approach for real-world scenarios where memory storage is not feasible. Building on this research, we hope to see further progress in exploring non-exemplar video continual learning methods in the future.

References

Alssum, L.; Alcazar, J. L.; Ramazanova, M.; Zhao, C.; and Ghanem, B. 2023. Just a Glimpse: Rethinking Temporal Information for Video Continual Learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2474–2483.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724–4733.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-Attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1290–1299.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N. C.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2016. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114: 3521 – 3526.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, 2556–2563. IEEE.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient Episodic Memory for Continual Learning. In *Neural Information Processing Systems*.

Park, J.; Kang, M.; and Han, B. 2021. Class-Incremental Learning for Action Recognition in Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13698–13707.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2016. iCaRL: Incremental Classifier and Representation Learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5533–5542.

Simonyan, K.; and Zisserman, A. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. *ArXiv*, abs/1406.2199.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Teja, V. P.; and Panda, P. 2020. Compression-aware Continual Learning using Singular Value Decomposition. *arXiv preprint arXiv:2009.01956*.

Thengane, V.; Khan, S.; Hayat, M.; and Khan, F. 2022. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*.

Van de Ven, G. M.; and Tolias, A. S. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Villa, A.; Alc'azar, J. L.; Alfarra, M.; Alhamoud, K.; Hurtado, J.; Heilbron, F. C.; Soto, Á.; and Ghanem, B. 2022a. PIVOT: Prompting for Video Continual Learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24214–24223.

Villa, A.; Alhamoud, K.; Alc'azar, J. L.; Heilbron, F. C.; Escorcia, V.; and Ghanem, B. 2022b. vCLIMB: A Novel Video Class Incremental Learning Benchmark. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19013–19022.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Gool, L. V. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision*.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022a. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, 631–648. Springer.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Yoon, J.; Yang, E.; Lee, J.; and Hwang, S. J. 2017. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*.

Zheng, Z.; Ma, M.; Wang, K.; Qin, Z.; Yue, X.; and You, Y. 2023. Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models. *arXiv preprint arXiv:2303.06628*.