Supervision	Approach	F1
Full	HMM + Text [24]	22.9%
	Discriminative Clustering[3]	41.4%
	KNN+GMM[30] + GT	**69.2%**
Weak	OCDC + Text Features [6]	28.9%
	OCDC [6]	**31.8%**
None	KNN+GMM[30]	32.2%
	Ours (RNN + No AL)	25.9%
	Ours (RNN + AL)	29.4%
	Ours (LSTM + No AL)	36.4%
	Ours (LSTM + AL)	**39.7%**

Table 3: Segmentation Results on the INRIA Instructional Videos dataset. We report F1 score for fair comparison.

We also evaluate the performance of the models with and without adaptive learning. It can be seen that long term temporal dependence captured by LSTMs is significant, especially due to the long durations of activities in the dataset. Additionally, the use of adaptive learning has a significant improvement in the segmentation framework, improving the performance by 9% and 11% for the RNN-based model and the LSTM-based model respectively, indicating a reduced overfitting of the model to the visual data.

4.4.1 Improved Features for Action Recognition

To evaluate the ability of the network to learn highly discriminative features for recognition, we evaluated the performance of the proposed approach in a recognition task. We use the model pretrained on the segmentation task on the Breakfast Actions dataset and use the hidden layer of the LSTM as input to a fully connected layer and use cross entropy to train the model. We also trained another network with the same structure - VGG16 + LSTM without the pretraining on the segmentation task to compare the effect of the features learned using self-supervision. As can be seen from Table 4, the use of self-supervision to pretrain the network prior to the recognition task improves the recognition performance of the network and has comparable performance to the other state-of-the-art approaches. It improves the recognition accuracy by 13.12% over the network without predictive pretraining.

Approach	Precision
HCF + HMM [19]	14.90%
HCF + CFG + HMM [19]	31.8%
RNN + ECTC [16]	35.6%
RNN + ECTC (Cosine) [16]	36.7%
HCF + Pattern Theory [9]	38.6%
HCF + Pattern Theory + ConceptNet[1]	42.9%
VGG16 + LSTM	33.54%
VGG16 + LSTM + Predictive Features(AL)	37.87%

Table 4: Activity recognition results on Breakfast Actions dataset. HCF and AL refer to handcrafted features and Adaptive Learning, respectively.

4.5. Qualitative Evaluation

Through the predictive, self supervised framework, we are able to learn the sequence of visual features in streaming video. We visualize the segmentation performance of the proposed framework on the Breakfast Actions Dataset in Figure 3. It can be seen that the proposed approach has high temporal coherence and does not suffer from over segmentation, especially when the segments are long. Long activity sequences allow the model to learn from observation by providing more samples of "*intra*-event" samples. Additionally, it can be seen that weakly supervised approaches like OCDC[6] and ECTC[16] suffer from over segmentation and intra-class fragmentation. This could arguably be attributed to the fact that they tend to enforce semantics, in the form of weak ordering of activities in the video regardless of the changes in visual features. Fully supervised approaches, such as HTK[20] perform better, especially due to the ability to assign semantics to visual features. However, they are also affected by unbalanced data and dataset shift, as can be seen in Figure 3 where the background class was segmented into other classes.

We also qualitatively evaluated the impact of adaptive learning and long term temporal memory in Figure 4, where the performance of the alternative methods described in Section 4.4. It can be seen that the use of adaptive learning during training allows the model to not overfit to any single class' intra-event frames and help generalize to other classes regardless of amount of training data. It is not to say that the problem of unbalanced data is alleviated, but the adaptive learning *does* help to some extent. It is interesting to note that the LSTM model tends to over-segment when not trained with adaptive learning, while the RNN-based model does not suffer from the same fate.

5. Conclusion

We demonstrate how a self-supervised learning paradigm can be used to segment long, highly complex visual sequences. There are key differences between our approach and fully supervised or weakly supervised approaches, including classical ones such as DBMs and HMMs. At a high level, our approach is unsupervised and does not require labeled data for training. The predictive error serves as supervision for training the framework. The other major aspect is that our approach requires only a single pass through the training data. Hence, the training time is very low. The experimental results demonstrate the robustness, high performance, and the generality of the approach on multiple real world datasets.