Action Localization through Continual Predictive Learning

Sathyanarayanan Aakur[1][0000−0003−1062−8929] and Sudeep Sarkar[2][0000−0001−7332−4207]

[1] Oklahoma State University, Stillwater, OK 74074 saakur@okstate.edu
[2] University of South Florida, Tampa, FL, 33620 sarkar@usf.edu

Abstract. The problem of action localization involves locating the action in the video, both over time and spatially in the image. The current dominant approaches use supervised learning to solve this problem. They require large amounts of annotated training data, in the form of frame-level bounding box annotations around the region of interest. In this paper, we present a new approach based on continual learning that uses feature-level predictions for self-supervision. It does not require any training annotations in terms of frame-level bounding boxes. The approach is inspired by cognitive models of visual event perception that propose a prediction-based approach to event understanding. We use a stack of LSTMs coupled with a CNN encoder, along with novel attention mechanisms, to model the events in the video and use this model to predict high-level features for the future frames. The prediction errors are used to learn the parameters of the models continuously. This self-supervised framework is not complicated as other approaches but is very effective in learning robust visual representations for both labeling and localization. It should be noted that the approach outputs in a streaming fashion, requiring only a single pass through the video, making it amenable for real-time processing. We demonstrate this on three datasets - UCF Sports, JHMDB, and THUMOS'13 and show that the proposed approach outperforms weakly-supervised and unsupervised baselines and obtains competitive performance compared to fully supervised baselines. Finally, we show that the proposed framework can generalize to egocentric videos and achieve state-of-the-art results on the *unsupervised* gaze prediction task. Code is available on the project page[3].

Keywords: Action localization, continuous learning, self-supervision

1 Introduction

We develop a framework for jointly learning spatial and temporal localization through continual, self-supervised learning, in a streaming fashion, requiring only *a single pass through the video*. Visual understanding tasks in computer vision have focused on the problem of recognition [1, 3, 23, 25] and captioning [1, 9, 47,

[3] https://saakur.github.io/Projects/ActionLocalization/

46], with the underlying assumption that each input video is already localized both spatially and temporally. While there has been tremendous progress in action localization, it has primarily been driven by the dependence on large amounts of tedious, spatial-temporal annotations. In this work, we aim to tackle the problem of spatial-temporal segmentation of streaming videos in a continual, self-supervised manner, without any training annotations.

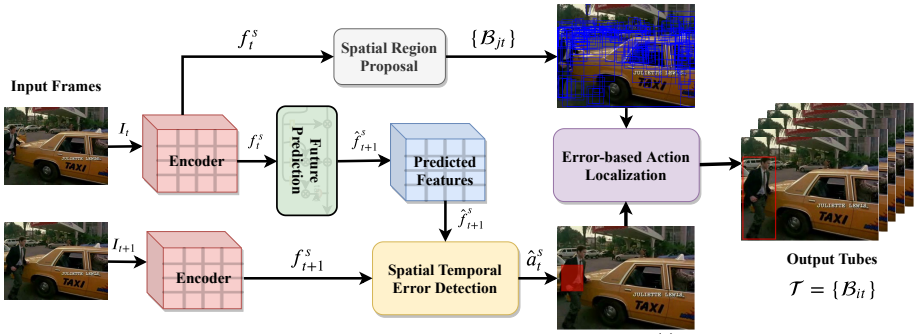


Fig. 1: The **Proposed Approach** has four core components: (i) feature extraction and spatial region proposal, (ii) a future prediction framework, (iii) a spatial-temporal error detection module and (iv) the error-based action localization process.

Drawing inspiration from psychology [13, 14, 52], we consider the underlying mechanism for both event understanding and attention selection in humans to be the idea of *predictability*. Defined as the surprise-attention hypothesis [13], unpredictable factors such as large changes in motion, appearance, or goals of the actor have a substantial effect on the event perception and human attention. Human event perception studies [52, 2] have shown that longer-term, temporal surprise have a strong correlation with event boundary detection. In contrast, short-term spatial surprise (such as those caused by motion) has a more substantial effect on human attention and localization [14]. Our approach combines both spatial and temporal surprise to formulate a computational framework to tackle the problem of self-supervised action localization in streaming videos in a continual manner.

We formulate our computational framework on the idea of spatial-temporal feature anticipation to model predictability of perceptual features. The main assumption in our framework is that expected, unpredictable features require attention and often point to the actor performing the action of interest. In contrast, predictable features can belong to background clutter and are not relevant to the action of interest. It is to be noted that unpredictability or *surprise* is not the same as *rarity*. It refers to short-term changes that aid in the completion of an overall task, which can be recurring [13]. We model the perceptual features using a hierarchical, cyclical, and recurrent framework, whose predictions are influenced by current and prior observations as well as current perceptual predictions. Hence, the predictive model's output can influence the perception of

the current frame being observed. The predictions are constantly compared with the incoming observations to provide self-supervision to guide future predictions.

We leverage these characteristics to derive and quantify spatial-temporal predictability. Our framework performs continuous learning to generate "*attention maps*" that overlap with the action being performed. Using these attention maps, we leverage advances in region proposals [29, 31, 44, 54] to localize actions in streaming videos without any supervision. Contrary to other attention-based approaches [5, 28, 33], we do not use the object-level characteristics such as label, role, and affordance in the proposal generation process.

Contributions: The contributions of our approach are three-fold: (i) we are among the first to tackle the problem of self-supervised action localization in *streaming videos without any training data such as labels or bounding boxes*, (ii) we show that modeling spatial-temporal prediction error can yield consistent localization performance across action classes and (iii) we show that the approach generalizes to egocentric videos and achieves competitive performance on the *unsupervised gaze prediction* task.

2 Related Work

Supervised action localization approaches tackle action localization through the simultaneous generation of bounding box proposals and labeling each bounding box with the predicted action class. Both bounding box generation and labeling are fully supervised, i.e., they require ground truth annotations of both bounding boxes and labels. Typical approachesleverage advances in object detection to include temporal information [7, 16, 18, 36, 37, 40, 43, 50] for proposal generation. The final step typically involves the use of the Viterbi algorithm [7] to link the generated bounding boxes across time.

Weakly-supervised action localization approaches have been explored to reduce the need for extensive annotations [5, 26, 28, 33]. They typically only require video-level labels and rely on object detection-based approaches to generate bounding box proposals. It is to be noted that weakly supervised approaches also use object-level labels and characteristics to guide the bounding box selection process. Some approaches [5] use a similarity-based tracker to connect bounding boxes across time to incorporate temporal consistency.

Unsupervised action localization approaches have not been explored to the same extent as supervised and weakly-supervised approaches. These approaches do not require any supervision - both labels or bounding boxes. The two more common approaches are to generate action proposals using (i) supervoxels [18, 38] and (ii) clustering motion trajectories [45]. It should be noted that [38] also uses object characteristics to evaluate the "humanness" of each super-voxel to select bounding box proposals. Our approach falls into the class of unsupervised action localization approaches. The most closely related approaches (with respect to architecture and theme) to ours are VideoLSTM [28] and Actor Supervision [5], which use attention in the selection process for gen-

erating bounding box proposals, but require video-level labels. We, on the other hand, do not require any labels or bounding box annotations for training.

While fully supervised approaches have more precise localization and achieve better recognition, the required number of annotations is rather large. It is not amenable to an increase in the number of classes and a decrease in the number of training videos. While not requiring frame-level annotations, weakly supervised approaches have the underlying assumption that there exists a large, annotated training set that allows for effective detection of all possible actors (both human and non-human) in the set of action classes. Unsupervised approaches, such as ours, do not make any such assumptions but can result in poorer localization performance. We alleviate this to an extent by leveraging advances in region proposal mechanisms and self-learning robust representations for obtaining video-level labels.

3 Self-Supervised Action Localization

In this section, we introduce our self-supervised action localization framework, as illustrated in Figure 1. Our approach has four core components: (i) feature extraction and spatial region proposal, (ii) a self-supervised future prediction framework, (iii) a spatial-temporal error detection module, and (iv) the error-based action localization process.

3.1 Feature Extraction and Spatial Region Proposal

The first step in our approach is feature extraction and the subsequent *per-frame* region proposal generation for identifying possible areas of actions and associated objects. Considering the tremendous advances in deep learning architectures for learning robust spatial representations, we use pre-trained convolutional neural networks to extract the spatial features for each frame in the video. We use a region proposal module, based on these spatial features, to predict possible action-agnostic spatial locations. We use class-agnostic proposals (i.e., the object category is ignored, and only feature-based localizations are taken into account) for two primary reasons. First, we do not want to make any assumptions on the actor's characteristics, such as label, role, and affordance. Second, despite significant progress in object detection, there can be many missed detections, especially when the object (or actor) performs actions that can transform their physical appearance. It is to be noted that these considerations can result in a large number of region proposals that require careful and robust selection but can yield higher chances of correct localization.

3.2 Self-supervised Future Prediction

The second stage in our proposed framework is the self-supervised future prediction framework. We consider the future prediction module to be a generative model whose output is conditioned on two factors - the current observation and

an *internal event model*. The current observation f_t^S is the feature-level encoding of the presently observed frame, I_t. We use the same feature encoder as the region proposal module to reduce the approach's memory footprint and complexity. The *internal event model* is a set of parameters that can effectively capture the spatial-temporal dynamics of the observed event. Formally, we define the predictor model as $P(\hat{f}_{t+1}^S|W_e, f_t^S)$, where W_e represents the internal event model and $\hat{f}_{t+1}^S$ is the predicted features at time $t+1$. Note that features f_t^S is not a one-dimensional vector, but a tensor (of dimension $w_f \times h_f \times d_f$) representing the features at each spatial location.

We model temporal dynamics of the observed event using Long Short Term Memory Networks (LSTMs)[12]. While other approaches [21, 48, 49] can be used for prediction, we consider LSTMs to be more suited for the following reasons. First, we want to model the temporal dynamics across *all* frames of the observed action (or event). Second, LSTMs can allow for multiple possible futures and hence will not tend to average the outcomes of these possible futures, as can be the case with other prediction models. Third, since we work with error-based localization, using LSTMs can ensure that the learning process propagates the spatial-temporal error across time and can yield progressively better predictions, especially for actions of longer duration. Formally, we can express LSTMs as

$$i_t = \sigma(W_i x_t + W_{hi} h_{t-1} + b_i); \quad f_t = \sigma(W_f x_t + W_{hf} h_{t-1} + b_f) \tag{1}$$

$$o_t = \sigma(W_o x_t + W_{ho} h_{t-1} + b_o); \quad g_t = \phi(W_g x_t + W_{hg} h_{t-1} + b_g) \tag{2}$$

$$m_t = f_t \cdot m_{t-1} + i_t \cdot g_t; \quad h_t = o_t \cdot \phi(m_t) \tag{3}$$

where x_t is the input at time t, σ is a non-linear activation function, $(\cdot)$ represents element-wise multiplication, ϕ is the hyperbolic tangent function ($tanh$) and W_k and b_k represent the trained weights and biases for each of the gates.

As opposed to [2], who also use an LSTM-based predictor and a decoder network, we use a hierarchical LSTM model (with three LSTM layers) as our event model. This modification allows us to model both spatial and temporal dependencies, since each higher-level LSTMs act as a progressive decoder framework that captures the temporal dependencies captured by the lower-level LSTMs. The first LSTM captures the spatial dependency that is propagated up the prediction stack. The updated hidden state of the first (bottom) LSTM layer (h_t^1) depends on the current observation f_t^S, the previous hidden state (h_{t-1}^1) and memory state (m_{t-1}^1). Each of the higher-level LSTMs at level l take the output of the bottom LSTM's output h_t^{l-1} and memory state m_t^{l-1} and can be defined as $(h_t^l, m_t^l) = LSTM(h_{t-1}^l, h_t^{l-1}, m_t^{l-1})$. Note this is different from a typical hierarchical LSTM model [35] in that the higher LSTMs are impacted by the output of the lower level LSTMs at current time step, as opposed to that from the previous time step. Collectively, the event model W_e is described by the learnable parameters and their respective biases from the hierarchical LSTM stack.

Hence, the top layer of the prediction stack acts as the decoder whose goal is to predict the next feature f_{t+1}^S given all previous predictions $\hat{f}_1^S, \hat{f}_2^S, \ldots \hat{f}_t^S$, an event model W_e and the current observation f_t^S. We model this prediction function as a log-linear model characterized by

$$\log\, p(\hat{f}^s_{t+1}|h^l_t) = \sum_{n=1}^{t} f(W_e, f^S_t) + log\, Z(h_t) \tag{4}$$

where h^l_t is the hidden state of the l^{th} level LSTM at time t and $Z(h_t)$ is a normalization constant. The LSTM prediction stack acts as a generative process for anticipating future features.

The **objective function** for training the predictive stack is a weighted zero order hold between the predicted features and the actual observed features, weighted by the zero order hold difference. The prediction error at time t is given by $E(t) = \frac{1}{n_f} \sum_{i=1}^{w_f} \sum_{j=1}^{h_f} e_{ij}$, where

$$e_{ij} = \hat{m}_t(i,j) \odot \|f^S_{t+1}(i,j) - \hat{f}^S_{t+1}(i,j)\|^2_{\ell_2} \tag{5}$$

Each feature f^S_t has dimensions $w_f \times h_f \times d_f$ and $\hat{m}_t(i,j)$ is a function that returns the zero order difference between the observed features at times t and $t+1$ at location (i,j). Note that the prediction is done at the feature level and not at the pixel level, so the spatial quantization is coarser than pixels.

3.3 Prediction Error-based Attention Map

At the core of our approach is the idea of spatial-temporal prediction error for localizing the actions of interest in the video. It takes into account the quality of the predictions made and the relative spatial alignment of the prediction errors. The input to the error detection module is the quantity from Equation 5. We compute a weight α_{ij} associated with each spatial location (i,j) in the predicted feature $\hat{f}^S_{t+1}$ as

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{w_k} \sum_{n=1}^{h_k} \exp(e_{mn})} \tag{6}$$

where e_{ij} represents the weighted prediction error at location (i,j) (Equation 5). It can be considered to be a function $a(f^S_t, h^l_{t-1})$ of the state of the top-most LSTM and the input feature f^S_t at time t. The resulting matrix is an error-based attention map that allows us to localize the prediction error at a specific spatial location. And the average spatial error over time, $E(t)$, is used for temporal localization.

One may remark that the formulation of α_{ij} is very similar to Bahdanau attention [4]. However, there are two key differences. First, our formulation is not parametrized and does not add to the number of learnable parameters in the framework. Second, our attention map is a characterization of the difficulty in anticipating unpredictable motion. In contrast, Bahdanau attention is an effort to increase the decoder's encoding ability and does not characterize the unpredictability of the future feature. We compare the use of both types of attention in Section 5.4, where we see that error-based localization is more suitable for our application.

3.4 Extraction of Action Tubes

The action localization module receives a stream of bounding box proposals and an error-based attention map to select an output tube. The action localization is a selection algorithm that filters *all* region proposals from Section 3.1 and returns the collection of proposals that have a higher probability of action localization. We do so by assigning an energy term to each of the bounding box proposals ($\mathcal{B}_{it}$) at time t and choosing the top k bounding boxes with least energy as our final proposals. The energy of a bounding box $\mathcal{B}_{it}$ is defined as

$$E(\mathcal{B}_{it}) = w_\alpha \ \phi(\alpha_{ij}, \mathcal{B}_{it}) + w_t \delta(\mathcal{B}_{it}, \{\mathcal{B}_{j,t-1}\}) \tag{7}$$

where $\phi(\cdot)$ is a function that returns a value characteristic of the distance between the bounding box center and location of maximum error, $\delta(\cdot)$ is a function that returns the minimum spatial distance between the current bounding box and the closest bounding box from the previous time step. The constants w_α and w_t are scaling factors. Note that $\delta(\cdot)$ is introduced to enforce temporal consistency in predictions, but we find that it is optional since the LSTM prediction stack implicitly enforces the temporal consistency through its memory states. In our experiments we set $k = 10, w_\alpha = 0.75$.

3.5 Implementation Details

In our experiments, we use a VGG-16 [34] network pre-trained on ImageNet as our feature extraction network. We use the output of the last convolutional layer before the fully connected layers as our spatial features. Hence the dimensions of the spatial features are $w_f = 14, h_f = 14, d_f = 512$. These output features are then used by an SSD [29] to generate bounding box proposals. Note that we take the generated bounding box proposals without taking into account classes and associated probabilities. We use a three-layer hierarchical LSTM model with the hidden state size as 512 as our predictor module. We use the vanilla LSTM as proposed in [12]. Video level-features are obtained by max-pooling the element-wise dot-product of the hidden state of the top-most LSTM and the attention values across time. We train with the adaptive learning mechanism proposed in [2], with the initial learning rate set to be 1×10^{-8} and scaling factors Δ_t^- and Δ_t^+ as 1×10^{-2} and 1×10^{-3}, respectively. The network was trained for 1 epoch on a computer with one Titan X Pascal.

4 Experimental Setup

4.1 Data

We evaluate our approach on three publicly available datasets for evaluating the proposed approach on the action localization task.

UCF Sports [32] is an action localization dataset consisting of 10 classes of sports actions such as skating and lifting collected from sports broadcasts. It

is an interesting dataset since it has a high concentration of distinct scenes and motions that make it challenging for localization and recognition. We use the splits (103 training and 47 testing videos) as defined in [26] for evaluation.

JHMDB [19] is composed of 21 action classes and 928 trimmed videos. All videos are annotated with human-joints for every frame. The ground truth bounding box for the action localization task is chosen such that the box encompasses all the joints. This dataset offers several challenges, such as increasing amounts of background clutter, high inter-class similarity, complex motion (including camera motion), and occluded objects of interest. We report all results as the average across all three splits.

THUMOS'13 [22] is a subset of the UCF-101 [39] dataset, consisting of 24 classes and 3,207 videos. Ground truth bounding boxes are provided for each of the classes for the action localization task. It is also known as the **UCF-101-24** dataset. Following prior works [28, 38], we perform our experiments and report results on the first split.

We also evaluate the proposed approach's generalization ability on egocentric videos by evaluating it on the *unsupervised gaze prediction task*. There has been evidence from cognitive psychology that there is a strong correlation between gaze points and action localization [41]. Hence, the gaze prediction task would be a reasonable measure of the generalization to action localization in egocentric videos. We evaluate the performance on the **GTEA Gaze** [6] dataset, which consists of 17 sequences of tasks performed by 14 subjects, with each sequence lasting about 4 minutes. We use the official splits for the GTEA datasets as defined in prior works [6].

4.2 Metrics and Baselines

For the **action localization** task, we follow prior works [28, 38] and report the mean average precision (mAP) at various overlap thresholds, obtained by computing the Intersection Over Union (IoU) of the predicted and ground truth bounding boxes. We also evaluate the quality of bounding box proposals by measuring the average, per-frame IoU, and the bounding box *recall* at varying overlap ratios.

Since ours is an unsupervised approach, we obtain class labels by clustering the learned representations using the *k-means* algorithm. While more complicated clustering may yield better recognition results [38], the k-means approach allows us to evaluate the robustness of learned features. We evaluate our approach in two settings K_{gt} and K_{opt}, where the number of clusters is set to the number of ground truth action classes and an optimal number obtained through the elbow method [24], respectively. From our experiments, we observe that K_{opt} is three times the number of ground truth classes, which is not unreasonable and has been a working assumption in other deep learning-based clustering approaches [11]. Clusters are mapped to the ground truth clusters for evaluation using the Hungarian method, as done in prior unsupervised approaches [20, 51].

We also compare against other LSTM and attention-based approaches (Section 5.3) to the action localization problem for evaluating the effectiveness of the proposed training protocol.

For the **gaze prediction** task, we evaluate the approaches using **Area Under the Curve** (AUC), which measures the area under the curve on saliency maps for true positive versus false-positive rates under various threshold values. We also report the **Average Angular Error** (AAE), which measures the angular distance between the predicted and ground truth gaze positions. Since our model's output is a saliency map, AUC is a more appropriate metric compared to average angular error (AAE), which requires specific locations.

5 Quantitative Evaluation

In this section, we present the quantitative evaluation of our approach on two different tasks, namely action localization, and egocentric gaze prediction. For the action localization task, we evaluate our approach on two aspects - the quality of proposals and spatial-temporal localization.

5.1 Quality of Localization Proposals

We first evaluate the quality of our localization proposals by assuming perfect class prediction. This allows us to independently assess the quality of localization performed in a self-supervised manner. We present the results of the evaluation in Table 1 and compare against fully supervised, weakly supervised, and unsupervised baselines. As can be seen, we outperform many supervised and weakly supervised baselines. APT [45] achieves a higher localization score. However, it produces, on average, $1,500$ proposals per video, whereas our approach returns approximately 10 proposals. A large number of localization proposals per video can lead to higher recall and IoU but makes the localization task, i.e., action labeling per video harder and can affect the ability to generalize across domains. Also, it should be noted that our approach produces proposals in *streaming* fashion, as opposed to many of the other approaches, which produce action tubes based on motion computed across the entire video. This can make real-time action localization in streaming videos harder.

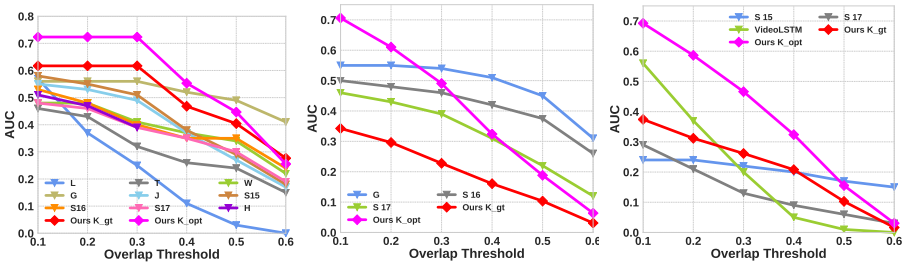
5.2 Spatial-temporal Action Localization

We also evaluate our approach on the spatial-temporal localization task. This evaluation allows us to analyze the robustness of the self-supervised features learned through prediction. We generate video-level class labels through clustering and use the standard evaluation metrics (Section 4.2) to quantify the performance. The AUC curves with respect to varying overlap thresholds are presented in Figure 2. We compare against a mix of supervised, weakly-supervised, and unsupervised baselines on all three datasets.

Supervision	Approach	Average
Full	STPD[42]	44.6
	Max Path Search [43]	**54.3**
Weak	Ma et al. [30]	44.6
	GBVS [8]	42.1
	Soomro et al. [38]	**47.7**
None	IME Tublets [18]	51.5
	APT [45]	**63.7**
	Proposed Approach	55.7

Table 1: Comparison with fully supervised and weakly supervised baselines on class-agnostic action localization on UCF Sports dataset. We report the average localization accuracy of each approach i.e. average IoU.

On the **UCF Sports** dataset (Figure 2(a)), we outperform all baselines including several supervised baselines except for Gkioxari and Malik [7] at higher overlap thresholds ($\sigma > 0.4$) when we set number of clusters k to the number of ground truth classes. When we allow for some over-segmentation and use the *optimal* number of clusters, we outperform all baselines till $\sigma > 0.5$.



(a) (b) (c)

Fig. 2: AUC for the action localization tasks are shown for (a) UCF Sports, (b) JHMDB and (c) THUMOS13 datasets. We compare against baselines with varying levels of supervision such as **L**an *et al.* [26], **T**ian *et al.* [40], **W**ang *et al.* [50], **G**kioxari and Malik [7], **J**ain *et al.* [18], **S**oomro *et al.* [36–38], **H**ou *et al.* [16], and VideoLSTM [28].

On the **JHMDB** dataset (Figure 2(b)), we find that our approach, while having high recall (77.8%@$\sigma = 0.5$), the large camera motion and intra-class variations have a significant impact on the classification accuracy. Hence, the mAP suffers when we set k to be the number of ground truth classes. When we set the number of clusters to the optimal number of clusters, we outperform other baselines at lower thresholds ($\sigma < 0.5$). It should be noted that the other unsupervised baseline (Soomro *et al.* [38]) uses object detection proposals from a Faster R-CNN backbone to score the "*humanness*" of a proposal. This assumption tends to make the approach biased towards human-centered action localization and affects its ability to generalize towards actions with non-human

actors. On the other hand, we do not make any assumptions on the characteristics of the actor, scene, or motion dynamics.

On the **THUMOS'13** dataset (Figure 2(c)), we achieve consistent improvements over unsupervised and weakly supervised baselines, at $k = k_{gt}$ and achieve state-of-the-art mAP scores when $k = k_{opt}$. It is interesting to note that we perform competitively (when $k = k_{gt}$) the weakly-supervised attention-based VideoLSTM [28], which uses a convLSTM for temporal modeling along with a CNN-based spatial attention mechanism. It should be noted that we have a higher recall rate ($0.47@\sigma = 0.4$ and $0.33@\sigma = 0.5$) at higher thresholds than other state-of-the-art approaches on THUMOS'13 and shows the robustness of the error-based localization approach to intra-class variation and occlusion.

Clustering quality. Since there is a significant difference in the mAP score when we set a different number of clusters in k-means, we measured the homogeneity (or purity) of the clustering. The homogeneity score measures the "quality" of the cluster by measuring how well a cluster models a given ground-truth class. Since we allow the over-segmentation of clusters when we set k to the optimal number of clusters, this is an essential measure of feature robustness. Higher homogeneity indicates that intra-class variations are captured since all data points in a given cluster belong to the same ground truth class. We observe an average homogeneity score of 74.56% when k is set to the number of ground truth classes and 78.97% when we use the optimal number of clusters. As can be seen, although we over-segment, each of the clusters typically models a single action class to a high degree of integrity.

Approach	Annotations		# Proposals	Average Recall					mAP
	Labels	Boxes		0.1	0.2	0.3	0.4	0.5	@0.2
ALSTM [33]	✓	✗	1	0.46	0.28	0.05	0.02	-	0.06
VideoLSTM [28]	✓	✗	1	0.71	0.52	0.32	0.11	-	0.37
Actor Supervision [5]	✓	✗	~ 1000	**0.89**	-	-	-	**0.44**	0.46
Proposed Approach	✗	✗	~ 10	0.84	**0.72**	**0.58**	**0.47**	0.33	**0.59**

Table 2: Comparison with other LSTM-based and attention-based approaches on the THUMOS'13 dataset. We report average recall at various overlap thresholds, mAP at 0.2 overlap threshold and the average number of proposals per frame.

5.3 Comparison with other LSTM-based approaches

We also compare our approach with other LSTM-based and attention-based models to highlight the importance of the proposed self-supervised learning paradigm. Since LSTM-based frameworks can have highly similar architectures, we consider different requirements and characteristics, such as the level of annotations required for training and the number of localization proposals returned per video. We compare with three approaches similar in spirit to our approach - ALSTM [33], VideoLSTM [28] and Actor Supervision [5] and summarize the results in Table 2. It can be seen that we significantly outperform VideoLSTM

and ALSTM on the THUMOS'13 dataset in both recall and $mAP@\sigma = 0.2$. Actor Supervision [5] outperforms our approach on recall, but it is to be noted that the region proposals are dependent on two factors - (i) object detection-based actor proposals and (ii) a filtering mechanism that limits proposals based on ground truth action classes, which can increase the training requirements and limit generalizability. Also, note that returning a higher number of localization proposals can increase recall at the cost of generalization.

5.4 Ablative Studies

The proposed approach has three major units that affect its performance the most - (i) the region proposal module, (ii) future prediction module, and (iii) error-based action localization module. We consider and evaluate several alternatives to all three modules.

We choose selective search [44] and EdgeBox [54] as alternative region proposal methods to SSD. We use an attention-based localization method for action localization as an approximation of the ALSTM [33] to evaluate the effectiveness of using the proposed error-based localization. We also evaluate a 1-layer LSTM predictor with a fully connected decoder network to approximate [2] on the localization task. We evaluate the effect of attention-based prediction by introducing a Bahdanau [4] attention layer before prediction as an alternative to the error-based action localization module.

These ablative studies are conducted on the UCF Sports dataset. The results are plotted in Figure 3(a). It can be seen that the use of the prediction error-based localization has a significant improvement over a trained attention-based localization approach. We can also see that the choice of region proposal methods do have some effect on the performance of the approach, with selective search and EdgeBox proposals doing slightly better at higher thresholds ($\sigma \in (0.4, 0.5)$) at the cost of inference time and additional bounding box proposals (50 compared to the 10 from SSD-based region proposal). Using SSD for generating proposals allows us to share weights across the frame encoder and region proposal tasks and hence reduce the memory and computational footprint of the approach. We also find that using attention as part of the prediction module significantly impacts the architecture's performance. It could, arguably, be attributed to the objective function, which aims to minimize the prediction error. Using attention to encode the input could impact the prediction function.

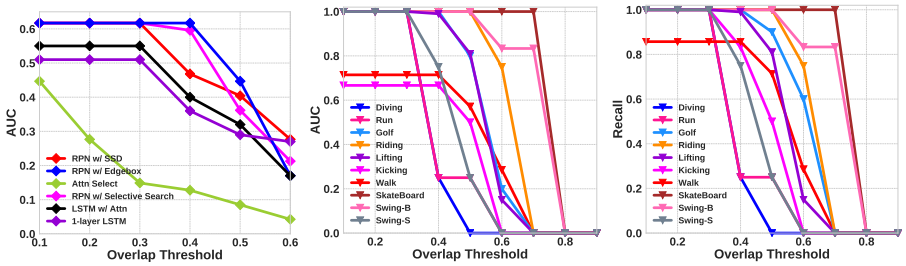
5.5 Unsupervised Egocentric Gaze Prediction

Finally, we evaluate the ability to generalize to egocentric videos by quantifying the model's performance on the unsupervised gaze prediction task. Given that we do not need any annotations or other auxiliary data, we employ the same architecture and training strategy for this task. We evaluate on the GTEA gaze dataset and compare it with other unsupervised models in Table 3. As can be seen, we obtain competitive results on the gaze prediction task, outperforming all baselines on both the AUC and AAE scores. It is to be noted that we outperform

	Itti *et al.* [17]	GBVS [10]	AWS-D [27]	Center Bias	OBDL [15]	Ours
AUC	0.747	0.769	0.770	0.789	0.801	**0.861**
AAE	18.4	15.3	18.2	**10.2**	15.6	13.6

Table 3: Comparison with state-of-the-art on the unsupervised egocentric gaze prediction task on the GTEA dataset.

the center bias method on the AUC metric. Center bias exploits the spatial bias in egocentric images and always predicts the center of the video frame as the predicted gaze position. The AUC metric's significant improvement indicates that our approach predicts gaze fixations that are more closely aligned with the ground truth than the center bias approach. Given that the model was not designed explicitly for this task, it is a remarkable performance, especially given the performance of fully supervised baselines such as DFG [53], which achieves 10.6 and 88.3 for AUC and AAE.



(a) (b) (c)

Fig. 3: Qualitative analysis of the proposed approach on UCF Sports dataset (a) ablative variations on AUC. (a) class-wise AUC, and (c) class-wise bounding box recall at different overlap thresholds.

5.6 Qualitative Evaluation

We find that our approach has a consistently high recall for the localization task across datasets and domains. We consider that an action is correctly localized if the average IoU across all frames is higher than 0.5, which indicates that most, if not all, frames in a video are correctly localized. We illustrate the recall scores and subsequent AUC scores for each class in the UCF sports dataset in Figures 3(b) and (c). For many classes (7/10 to be specific), we have more than 80% recall at an overlap threshold of 0.5. We find, through visual inspection, that the spatial-temporal error is often correlated with the actor, but is usually not at the center of the region of interest and thus reduces the quality of the chosen proposals. We illustrate this effect in Figure 4. The first row shows the input frame, the second shows the error-based attention, and the last row shows the final localization proposals. If more proposals are returned (as is the case with selective search and EdgeBox), we can obtain a higher recall (Figure 3(b)) and higher mAP.

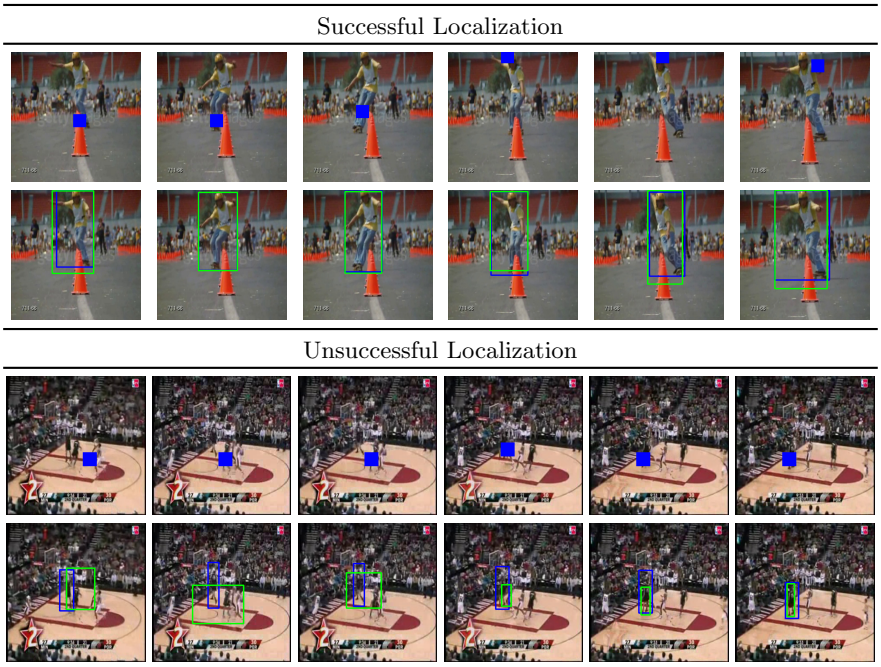


Fig. 4: **Qualitative Examples**: We present the error-based attention location and the final prediction, for both successful and unsuccessful localizations. Green BB: Prediction, Blue BB: Ground truth

6 Conclusion

In this work, we introduce a self-supervised approach to action localization, driven by spatial-temporal error localization. We show that the use of self-supervised prediction using video frames can help learn highly robust features and obtain state-of-the-art results on localization without any training annotations. We also show that the proposed framework can work with a variety of proposal generation methods without losing performance. We also show that the approach can generalize to egocentric videos without changing the training methodology or the framework and obtain competitive performance on the unsupervised gaze prediction task.

Acknowledgement

This research was supported in part by the US National Science Foundation grants CNS 1513126, IIS 1956050, and IIS 1955230.

References

1. Aakur, S., de Souza, F.D., Sarkar, S.: Going deeper with semantics: Exploiting semantic contextualization for interpretation of human activity in videos. In: IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2019)
2. Aakur, S.N., Sarkar, S.: A perceptual prediction framework for self supervised event segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
3. Aakur, S.N., de Souza, F.D., Sarkar, S.: Towards a knowledge-based approach for generating video descriptions. In: Conference on Computer and Robot Vision (CRV). Springer (2017)
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
5. Escorcia, V., Dao, C.D., Jain, M., Ghanem, B., Snoek, C.: Guess where? actor-supervision for spatiotemporal action localization. Computer Vision and Image Understanding **192**, 102886 (2020)
6. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: European Conference on Computer Vision. pp. 314–327. Springer (2012)
7. Gkioxari, G., Malik, J.: Finding action tubes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 759–768 (2015)
8. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: 2010 ieee computer society conference on computer vision and pattern recognition. pp. 2141–2148. IEEE (2010)
9. Guo, Z., Gao, L., Song, J., Xu, X., Shao, J., Shen, H.T.: Attention-based lstm with semantic consistency for videos captioning. In: ACM Conference on Multimedia (ACM MM). pp. 357–361. ACM (2016)
10. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems. pp. 545–552 (2007)
11. Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 31–35. IEEE (2016)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
13. Horstmann, G., Herwig, A.: Surprise attracts the eyes and binds the gaze. Psychonomic Bulletin & Review **22**(3), 743–749 (2015)
14. Horstmann, G., Herwig, A.: Novelty biases attention and gaze in a surprise trial. Attention, Perception, & Psychophysics **78**(1), 69–77 (2016)
15. Hossein Khatoonabadi, S., Vasconcelos, N., Bajic, I.V., Shan, Y.: How many bits does it take for a stimulus to be salient? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2015)
16. Hou, R., Chen, C., Shah, M.: Tube convolutional neural network (t-cnn) for action detection in videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5822–5831 (2017)
17. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research **40**(10-12), 1489–1506 (2000)
18. Jain, M., Van Gemert, J., Jégou, H., Bouthemy, P., Snoek, C.G.: Action localization with tubelets from motion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 740–747 (2014)

19. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 3192–3199 (2013)
20. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9865–9874 (2019)
21. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: Neural Information Processing Systems. pp. 667–675 (2016)
22. 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)
23. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1725–1732 (2014)
24. Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. International Journal **1**(6), 90–95 (2013)
25. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 780–787 (2014)
26. Lan, T., Wang, Y., Mori, G.: Discriminative figure-centric models for joint action localization and recognition. In: 2011 International conference on computer vision. pp. 2003–2010. IEEE (2011)
27. Leboran, V., Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M.: Dynamic whitening saliency. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(5), 893–907 (2016)
28. Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.: Videolstm convolves, attends and flows for action recognition. Computer Vision and Image Understanding **166**, 41–50 (2018)
29. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
30. Ma, S., Zhang, J., Ikizler-Cinbis, N., Sclaroff, S.: Action recognition and localization by hierarchical space-time segments. In: Proceedings of the IEEE international conference on computer vision. pp. 2744–2751 (2013)
31. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)
32. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2008)
33. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. In: Neural Information Processing Systems: Time Series Workshop (2015)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
35. Song, J., Gao, L., Guo, Z., Liu, W., Zhang, D., Shen, H.T.: Hierarchical lstm with adjusted temporal attention for video captioning. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 2737–2743. AAAI Press (2017)
36. Soomro, K., Idrees, H., Shah, M.: Action localization in videos through context walk. In: Proceedings of the IEEE international conference on computer vision. pp. 3280–3288 (2015)

37. Soomro, K., Idrees, H., Shah, M.: Predicting the where and what of actors and actions through online action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2648–2657 (2016)
38. Soomro, K., Shah, M.: Unsupervised action discovery and localization in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 696–705 (2017)
39. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
40. Tian, Y., Sukthankar, R., Shah, M.: Spatiotemporal deformable part models for action detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2642–2649 (2013)
41. Tipper, S.P., Lortie, C., Baylis, G.C.: Selective reaching: Evidence for action-centered attention. Journal of Experimental Psychology: Human Perception and Performance **18**(4), 891 (1992)
42. Tran, D., Yuan, J.: Optimal spatio-temporal path discovery for video event detection. In: CVPR 2011. pp. 3321–3328. IEEE (2011)
43. Tran, D., Yuan, J.: Max-margin structured output regression for spatio-temporal action localization. In: Advances in neural information processing systems. pp. 350–358 (2012)
44. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International Journal of Computer Vision (IJCV) **104**(2), 154–171 (2013)
45. Van Gemert, J.C., Jain, M., Gati, E., Snoek, C.G., et al.: Apt: Action localization proposals from dense trajectories. In: BMVC. vol. 2, p. 4 (2015)
46. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: IEEE International Conference on Computer Vision (ICCV). pp. 4534–4542 (2015)
47. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014)
48. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 98–106 (2016)
49. Vondrick, C., Torralba, A.: Generating the future with adversarial transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1020–1028 (2017)
50. Wang, L., Qiao, Y., Tang, X.: Video action detection with relational dynamic-poselets. In: European conference on computer vision. pp. 565–580. Springer (2014)
51. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning (ICML). pp. 478–487 (2016)
52. Zacks, J.M., Tversky, B., Iyer, G.: Perceiving, remembering, and communicating structure in events. Journal of Experimental Psychology: General **130**(1), 29 (2001)
53. Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., Feng, J.: Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4372–4381 (2017)
54. Zhu, G., Porikli, F., Li, H.: Tracking randomly moving objects on edge box proposals. arXiv preprint arXiv:1507.08085 (2015)