In this paper, we introduce a framework for improving object detection in videos by capturing temporal context and encouraging temporally consistent predictions. First, we train a pseudo-labeler, that is, a domain-adapted convolutional neural network for object detection. The pseudo-labeler is first trained individually on the subset of labeled frames, and then subsequently applied to all frames. Then we train a recurrent neural network (RNN) that takes as input sequences of pseudo-labeled frames and optimizes an objective that encourages both accuracy on the target frame and consistency across consecutive frames.

The approach incorporates strong supervision of target frames, weak-supervision on context frames, and regularization via a smoothness penalty. Building on YOLO, a domain-adapted frame-level object detection model [3], we demonstrate that for the sparsely annotated YouTube Objects dataset [2], our method achieves mean Average Precision (mAP) of 68.73 on test data, as compared to a best published result of 37.41 [4] and 61.66 for YOLO alone.

As with YOLO [3], our fine-tuned pseudo-labeler takes $448 \times 448$ frames as input and progresses on category types and locations of possible objects at each one of $7 \times 7$ non-overlapping grid cells. For each grid cell, the model outputs class conditional probabilities as well as 2 bounding boxes and their associated confidence scores.

Then, to incorporate temporal context, we train an RNN with gated recurrent units (GRUs) [1] to refine the provisional predictions. This net takes as input sequences of pseudo-labels. For this recurrent model, we demonstrate an efficient and effective training strategy. The objective encourages predictions to be close to true labels (for labeled frames), not to deviate too far from the pseudo-labels, and to be similar across adjacent frames. As demonstrated experimentally, our framework proves effective, achieving state-of-the-art mAP and producing compelling visual examples.