Context Matters: Refining Object Detection in Video with Recurrent Neural Networks

Subarna Tripathi[1]
stripathi@ucsd.edu
Zachary C. Lipton[1]
zlipton@cs.ucsd.edu
Serge Belongie[2, 3]
sjb344@cornell.edu
Truong Nguyen[1]
tqn001@eng.ucsd.edu

[1] University of California San Diego
La Jolla, CA, USA
[2] Cornell University
Ithaca, NY, USA
[3] Cornell Tech
New York, NY, USA

Abstract

Given the vast amounts of video available online, and recent breakthroughs in object detection with static images, object detection in video offers a promising new frontier. However, motion blur and compression artifacts cause substantial frame-level variability, even in videos that appear smooth to the eye. Additionally, video datasets tend to have sparsely annotated frames. We present a new framework for improving object detection in videos that captures temporal context and encourages consistency of 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 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. Our approach achieves mean Average Precision (mAP) of 68.73, an improvement of 7.1 over the strongest image-based baselines for the Youtube-Video Objects dataset. Our experiments demonstrate that neighboring frames can provide valuable information, even absent labels.

1 Introduction

Despite the immense popularity and availability of online video content via outlets such as Youtube and Facebook, most work on object detection focuses on static images. Given the breakthroughs of deep convolutional neural networks for detecting objects in static images, the application of these methods to video might seem straightforward. However, motion blur and compression artifacts cause substantial frame-to-frame variability, even in videos that appear smooth to the eye. These attributes complicate prediction tasks like classification and localization. Object-detection models trained on images tend not to perform competitively on videos owing to domain shift factors [12]. Moreover, object-level annotations in popular

video data-sets can be extremely sparse, impeding the development of better video-based object detection models.

Girshik *et al*. [9] demonstrate that even given scarce labeled training data, high-capacity convolutional neural networks can achieve state of the art detection performance if first pre-trained on a related task with abundant training data, such as 1000-way ImageNet classification. Followed the pretraining, the networks can be fine-tuned to a related but distinct domain. Also relevant to our work, the recently introduced models Faster R-CNN [21] and You Look Only Once (YOLO) [20] unify the tasks of classification and localization. These methods, which are accurate and efficient, propose to solve both tasks through a single model, bypassing the separate object proposal methods used by R-CNN [9].

In this paper, we introduce a method to extend unified object recognition and localization to the video domain. Our approach applies transfer learning from the image domain to video frames. Additionally, we present a novel recurrent neural network (RNN) method that refines predictions by exploiting contextual information in neighboring frames. In summary, we contribute the following:

- A new method for refining a video-based object detection consisting of two parts: (i) a *pseudo-labeler*, which assigns provisional labels to all available video frames. (ii) A recurrent neural network, which reads in a sequence of provisionally labeled frames, using the contextual information to output refined predictions.

- An effective training strategy utilizing (i) category-level weak-supervision at every time-step, (ii) localization-level strong supervision at final time-step (iii) a penalty encouraging prediction smoothness at consecutive time-steps, and (iv) similarity constraints between *pseudo-labels* and prediction output at every time-step.

- An extensive empirical investigation demonstrating that on the YouTube Objects [19] dataset, our framework achieves mean average precision (mAP) of 68.73 on test data, compared to a best published result of 37.41 [26] and 61.66 for a domain adapted YOLO network [20].

2 Methods

In this work, we aim to refine object detection in video by utilizing contextual information from neighboring video frames. We accomplish this through a two-stage process. First, we train a *pseudo-labeler*, that is, a domain-adapted convolutional neural network for object detection, trained individually on the labeled video frames. Specifically, we fine-tune the YOLO object detection network [20], which was originally trained for the 20-class PASCAL VOC [5] dataset, to the Youtube-Video [19] dataset.

When fine-tuning to the 10 sub-categories present in the video dataset, our objective is to minimize the weighted squared detection loss (equation 3) as specified in YOLO [20]. While fine-tuning, we learn only the parameters of the top-most fully-connected layers, keeping the 24 convolutional layers and 4 max-pooling layers unchanged. The training takes roughly 50 epochs to converge, using the RMSProp [25] optimizer with momentum of 0.9 and a mini-batch size of 128.

As with YOLO [20], our fine-tuned $pseudo-labeler$ takes 448×448 frames as input and regresses on category types and locations of possible objects at each one of $S \times S$ non-overlapping grid cells. For each grid cell, the model outputs class conditional probabilities

as well as B bounding boxes and their associated confidence scores. As in YOLO, we consider a *responsible* bounding box for a grid cell to be the one among the B boxes for which the predicted area and the ground truth area shares the maximum Intersection Over Union. During training, we simultaneously optimize classification and localization error (equation 3). For each grid cell, we minimize the localization error for the *responsible* bounding box with respect to the ground truth only when an object appears in that cell.

Next, we train a Recurrent Neural Network (RNN), with Gated Recurrent Units (GRUs) [2]. This net takes as input sequences of *pseudo-labels*, optimizing an objective that encourages both accuracy on the target frame and consistency across consecutive frames. Given a series of *pseudo-labels* $\mathbf{x}^{(1)},...,\mathbf{x}^{(T)}$, we train the RNN to generate improved predictions $\hat{\mathbf{y}}^{(1)},...,\hat{\mathbf{y}}^{(T)}$ with respect to the ground truth $\mathbf{y}^{(T)}$ available only at the final step in each sequence. Here, t indexes sequence steps and T denotes the length of the sequence. As output, we use a fully-connected layer with a linear activation function, as our problem is regression. In our final experiments, we use a 2-layer GRU with 150 nodes per layer, hyper-parameters determined on validation data.

The following equations define the forward pass through a GRU layer, where $\mathbf{h}_l^{(t)}$ denotes the layer's output at the current time step, and $\mathbf{h}_{l-1}^{(t)}$ denotes the previous layer's output at the same sequence step:

$$\begin{aligned}
\mathbf{r}_l^{(t)} &= \sigma(\mathbf{h}_{l-1}^{(t)} W_l^{xr} + \mathbf{h}_l^{(t-1)} W_l^{hr} + \mathbf{b}_l^r) \\
\mathbf{u}_l^{(t)} &= \sigma(\mathbf{h}_{l-1}^{(t)} W_l^{xu} + \mathbf{h}_l^{(t-1)} W_l^{hu} + \mathbf{b}_l^u) \\
\mathbf{c}_l^{(t)} &= \sigma(\mathbf{h}_{l-1}^{(t)} W_l^{xc} + r_t \odot (\mathbf{h}_l^{(t-1)} W_l^{hc}) + \mathbf{b}_l^c) \\
\mathbf{h}_l^{(t)} &= (1 - \mathbf{u}_l^{(t)}) \odot \mathbf{h}_l^{(t-1)} + \mathbf{u}_l^{(t)} \odot \mathbf{c}_l^{(t)}
\end{aligned} \tag{1}$$

Here, σ denotes an element-wise logistic function and $\odot$ is the (element-wise) Hadamard product. The reset gate, update gate, and candidate hidden state are denoted by $\mathbf{r}$, $\mathbf{u}$, and $\mathbf{c}$ respectively. For $S = 7$ and $B = 2$, the pseudo-labels $\mathbf{x}^{(t)}$ and prediction $\hat{\mathbf{y}}^{(t)}$ both lie in $\mathbb{R}^{1470}$.

2.1 Training

We design an objective function (Equation 2) that accounts for both accuracy at the target frame and consistency of predictions across adjacent time steps in the following ways:

$$\text{loss} = \text{d\_loss} + \alpha \cdot \text{s\_loss} + \beta \cdot \text{c\_loss} + \gamma \cdot \text{pc\_loss} \tag{2}$$

Here, d_loss, s_loss, c_loss and pc_loss stand for detection_loss, similarity_loss, category_loss and prediction_consistency_loss described in the following sections. The values of the hyper-parameters $\alpha = 0.2$, $\beta = 0.2$ and $\gamma = 0.1$ are chosen based on the detection performance on the validation set. The training converges in 80 epochs for parameter updates using RMSProp [25] and momentum 0.9. During training we use a mini-batch size of 128 and sequences of length 30.

2.1.1 Strong Supervision at Target Frame

On the final output, for which the ground truth classification and localization is available, we apply a multi-part object detection loss as described in YOLO [20].

$$\begin{aligned}\text{detection\_loss} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left(x_i^{(T)} - \hat{x}_i^{(T)}\right)^2 + \left(y_i^{(T)} - \hat{y}_i^{(T)}\right)^2 \\ + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left(\sqrt{w_i}^{(T)} - \sqrt{\hat{w}_i^{(T)}}\right)^2 + \left(\sqrt{h_i}^{(T)} - \sqrt{\hat{h}_i^{(T)}}\right)^2 \\ + \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} (C_i - \hat{C}_i)^2 \\ + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj} \left(C_i^{(T)} - \hat{C}_i^{(T)}\right)^2 \\ + \sum_{i=0}^{S^2} \mathbb{1}_{i}^{obj} \sum_{c \in classes} \left(p_i^{(T)}(c) - \hat{p}_i{}^{(T)}(c)\right)^2\end{aligned} \tag{3}$$

where $\mathbb{1}_i^{obj}$ denotes if the object appears in cell i and $\mathbb{1}_{ij}^{obj}$ denotes that jth bounding box predictor in cell i is *responsible* for that prediction. The loss function penalizes classification and localization error differently based on presence or absence of an object in that grid cell. x_i, y_i, w_i, h_i corresponds to the ground truth bounding box center coordinates, width and height for objects in grid cell (if it exists) and $\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i$ stand for the corresponding predictions. C_i and $\hat{C}_i$ denote confidence score of *objectness* at grid cell i for ground truth and prediction. $p_i(c)$ and $\hat{p}_i(c)$ stand for conditional probability for object class c at cell index i for ground truth and prediction respectively. We use similar settings for YOLO's object detection loss minimization and use values of $\lambda_{coord} = 5$ and $\lambda_{noobj} = 0.5$.

2.1.2 Similarity between *Pseudo-labels* and Predictions

Our objective function also includes a regularizer that penalizes the dissimilarity between *pseudo-labels* and the prediction at each time frame t.

$$\text{similarity\_loss} = \sum_{t=0}^{T} \sum_{i=0}^{S^2} \hat{C}_i^{(t)} \left(\mathbf{x}_i^{(t)} - \hat{\mathbf{y}}_i{}^{(t)}\right)^2 \tag{4}$$

Here, $\mathbf{x}_i^{(t)}$ and $\hat{\mathbf{y}}_i{}^{(t)}$ denote the *pseudo-labels* and predictions corresponding to the i-th grid cell at t-th time step respectively. We perform minimization of the square loss weighted by the predicted confidence score at the corresponding cell.

2.1.3 Object Category-level Weak-Supervision

Replication of the static target at each sequential step has been shown to be effective in [3, 16, 28]. Of course, with video data, different objects may move in different directions and speeds. Yet, within a short time duration, we could expect all objects to be present. Thus we employ target replication for classification but not localization objectives.

We minimize the square loss between the categories aggregated over all grid cells in the ground truth $\mathbf{y}^{(T)}$ at final time step T and predictions $\hat{\mathbf{y}}^{(t)}$ at all time steps t. Aggregated category from the ground truth considers only the cell indices where an object is present. For predictions, contribution of cell i is weighted by its predicted confidence score $\hat{C}_i^{(t)}$. Note that cell indices with positive detection are sparse. Thus, we consider the confidence score of each cell while minimizing the aggregated category loss.

$$\text{category\_loss} = \sum_{t=0}^{T} \left(\sum_{c \in classes} \left(\sum_{i=0}^{S^2} \hat{C}_i^{(t)} \left(\hat{p}_i^{(t)}(c) \right) - \sum_{i=0}^{S^2} \mathbb{1}_i^{obj^{(T)}} \left(p_i^{(T)}(c) \right) \right) \right)^2 \tag{5}$$

2.1.4 Consecutive Prediction Smoothness

Additionally, we regularize the model by encouraging smoothness of predictions across consecutive time-steps. This makes sense intuitively because we assume that objects rarely move rapidly from one frame to another.

$$\text{prediction\_consistency\_loss} = \sum_{t=0}^{T-1} \left(\hat{\mathbf{y}}_i{}^{(t)} - \hat{\mathbf{y}}_i{}^{(t+1)} \right)^2 \tag{6}$$

2.2 Inference

The recurrent neural network predicts output at every time-step. The network predicts 98 bounding boxes per video frame and class probabilities for each of the 49 grid cells. We note that for every cell, the net predicts class conditional probabilities for each one of the C categories and B bounding boxes. Each one of the B predicted bounding boxes per cell has an associated *objectness* confidence score. The predicted confidence score at that grid is the maximum among the boxes. The bounding box with the highest score becomes the *responsible* prediction for that grid cell i.

The product of class conditional probability $\hat{p}_i^{(t)}(c)$ for category type c and *objectness* confidence score $\hat{C}_i^{(t)}$ at grid cell i, if above a threshold, infers a detection. In order for an object of category type c to be detected for i-th cell at time-step t, both the class conditional probability $\hat{p}_i^{(t)}(c)$ and *objectness score* $\hat{C}_i^{(t)}$ must be reasonably high.

Additionally, we employ Non-Maximum Suppression (NMS) to winnow multiple high scoring bounding boxes around an object instance and produce a single detection for an instance. By virtue of YOLO-style prediction, NMS is not critical.

3 Experimental Results

In this section, we empirically evaluate our model on the popular **Youtube-Objects** dataset, providing both quantitative results (as measured by mean Average Precision) and subjective evaluations of the model's performance, considering both successful predictions and failure cases.

The **Youtube-Objects** dataset[19] is composed of videos collected from Youtube by querying for the names of 10 object classes of the PASCAL VOC Challenge. It contains 155 videos in total and between 9 and 24 videos for each class. The duration of each video varies between 30 seconds and 3 minutes. However, only 6087 frames are annotated with 6975 bounding-box instances. The training and test split is provided.

3.1 Experimental Setup

We implement the domain-adaption of YOLO and the proposed RNN model using Theano [24]. Our best performing RNN model uses two GRU layers of 150 hidden units each and dropout of probability 0.5 between layers, significantly outperforming domain-adapted YOLO alone. While we can only objectively evaluate prediction quality on the labeled frames, we present subjective evaluations on sequences.

Average Precision on 10-categories

Methods	airplane	bird	boat	car	cat	cow	dog	horse	mbike	train
DPM[6]	28.42	48.14	25.50	48.99	1.69	19.24	15.84	35.10	31.61	39.58
VOP[26]	29.77	28.82	35.34	41.00	33.7	57.56	34.42	54.52	29.77	29.23
YOLO[20]	76.67	89.51	57.66	65.52	43.03	53.48	55.81	36.96	24.62	62.03
DA YOLO	**83.89**	**91.98**	59.91	81.95	46.67	56.78	53.49	42.53	32.31	67.09
RNN-IOS	82.78	89.51	68.02	**82.67**	47.88	70.33	52.33	61.52	27.69	**67.72**
RNN-WS	77.78	89.51	**69.40**	78.16	51.52	**78.39**	47.09	81.52	36.92	62.03
RNN-PS	76.11	87.65	62.16	80.69	**62.42**	78.02	**58.72**	**81.77**	**41.54**	58.23

Table 1: Per-category object detection results for the Deformable Parts Model (DPM), Video Object Proposal based AlexNet (VOP), image-trained YOLO (YOLO), domain-adapted YOLO (DA-YOLO). RNN-IOS regularizes on input-output similarity, to which RNN-WS adds category-level weak-supervision, to which RNN-PS adds a regularizer encouraging prediction smoothness.

3.2 Objective Evaluation

We compare our approach with other methods evaluated on the Youtube-Objects dataset. As shown in Table 3.2 and Table 3.2, Deformable Parts Model (DPM) [6])-based detector reports [12] mean average precision below 30, with especially poor performance in some categories such as *cat*. The method of Tripathi *et al.* (VPO) [26] uses consistent video object proposals followed by a domain-adapted AlexNet classifier (5 convolutional layer, 3 fully connected) [14] in an R-CNN [9]-like framework, achieving mAP of 37.41. We also compare against YOLO (24 convolutional layers, 2 fully connected layers), which unifies the classification and localization tasks, and achieves mean Average Precision over 55.

In our method, we adapt YOLO to generate *pseudo-labels* for all video frames, feeding them as inputs to the refinement RNN. We choose YOLO as the *pseudo-labeler* because it is the most accurate among feasibly fast image-level detectors. The domain-adaptation improves YOLO's performance, achieving mAP of 61.66.

Our model with RNN-based prediction refinement, achieves superior aggregate mAP to all baselines. The RNN refinement model using both input-output similarity, category-level weak-supervision, and prediction smoothness performs best, achieving 68.73 mAP. This amounts to a relative improvement of 11.5% over the best baselines. Additionally, the RNN improves detection accuracy on most individual categories (Table 3.2).

mean Average Precision on all categories

Methods	DPM	VOP	YOLO	DA YOLO	RNN-IOS	RNN-WS	RNN-PS
mAP	29.41	37.41	56.53	**61.66**	65.04	67.23	**68.73**

Table 2: Overall detection results on Youtube-Objects dataset. Our best model (RNN-PS) provides 7% improvements over DA-YOLO baseline.

3.3 Subjective Evaluation

We provide a subjective evaluation of the proposed RNN model in Figure 1. Top and bottom rows in every pair of sequences correspond to *pseudo-labels* and results from our approach respectively. While only the last frame in each sequence has associated ground truth, we can observe that the RNN produces more accurate and more consistent predictions across

time frames. The predictions are consistent with respect to classification, localization and confidence scores.

In the first example, the RNN consistently detects the *dog* throughout the sequence, even though the *pseudo-labels* for the first two frames were wrong (*bird*). In the second example, *pseudo-labels* were *motorbike*, *person*, *bicycle* and even *none* at different time-steps. However, our approach consistently predicted *motorbike*. The third example shows that the RNN consistently predicts both of the cars while the *pseudo-labeler* detects only the smaller car in two frames within the sequence. The last two examples show how the RNN increases its confidence scores, bringing out the positive detection for *cat* and *car* respectively both of which fell below the detection threshold of the *pseudo-labeler*.

3.4 Areas For Improvement

The YOLO scheme for unifying classification and localization [20] imposes strong spatial constraints on bounding box predictions since each grid cell can have only one class. This restricts the set of possible predictions, which may be undesirable in the case where many objects are in close proximity. Additionally, the rigidity of the YOLO model may present problems for the refinement RNN, which encourages smoothness of predictions across the sequence of frames. Consider, for example, an object which moves slightly but transits from one grid cell to another. Here smoothness of predictions seems undesirable.

Figure 2 shows some failure cases. In the first case, the *pseudo-labeler* classifies the instances as *dogs* and even as *birds* in two frames whereas the ground truth instances are *horses*. The RNN cannot recover from the incorrect pseudo-labels. Strangely, the model increases the confidence score marginally for a different wrong category *cow*. In the second case, possibly owing to motion and close proximity of multiple instances of the same object category, the RNN predicts the correct category but fails on localization. These point to future work to make the framework robust to motion.

The category-level weak supervision in the current scheme assumes the presence of all objects in nearby frames. While for short snippets of video this assumption generally holds, it may be violated in case of occlusions, or sudden arrival or departure of objects. In addition, our assumptions regarding the desirability of prediction smoothness can be violated in the case of rapidly moving objects.

4 Related Work

Our work builds upon a rich literature in both image-level object detection,video analysis, and recurrent neural networks. Several papers propose ways of using deep convolutional networks for detecting objects [1, 8, 9, 10, 17, 20, 21, 22, 23, 27]. Some approaches classify the proposal regions [9, 10] into object categories and some other recent methods [20, 21] unify the localization and classification stages. Kalogeiton *et al*. [12] identifies domain shift factors between still images and videos, necessitating video-specific object detectors. To deal with shift factors and sparse object-level annotations in video, researchers have proposed several strategies. Recently, [26] proposed both transfer learning from the image domain to video frames and optimizing for temporally consistent object proposals. Their approach is capable of detecting both moving and static objects. However, the object proposal generation step that precedes classification is slow.

Prest *et al*. [18], utilize weak supervision for object detection in videos via category-level annotations of frames, absent localization ground truth. This method assumes that the target

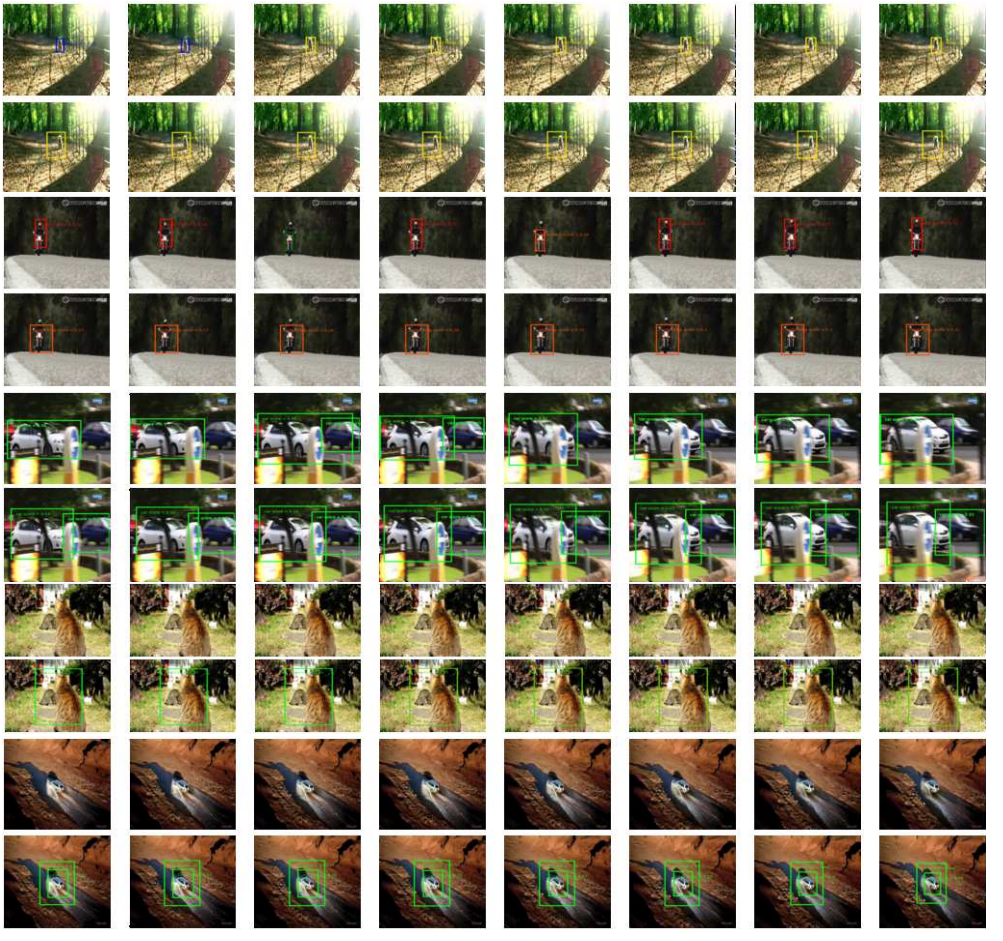

Figure 1: Object detection results from the final eight frames of five different test-set sequences. In each pair of rows, the top row shows the *pseudo-labeler* and the bottom row shows the RNN. In the first two examples, the RNN consistently predicts correct categories *dog* and *motorbike*, in contrast to the inconsistent baseline. In the third sequence, the RNN correctly predicts multiple instances while the *pseudo-labeler* misses one. For the last two sequences, the RNN increases the confidence score, detecting objects missed by the baseline.

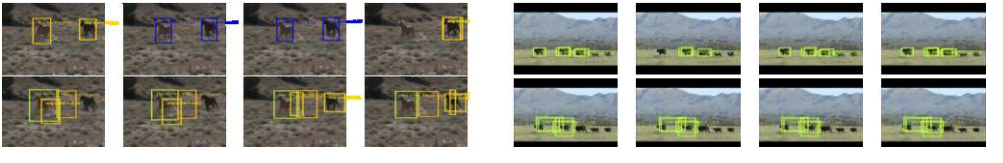

Figure 2: Failure cases for the proposed model. Left: the RNN cannot recover from incorrect *pseudo-labels*. Right: RNN localization performs worse than *pseudo-labels* possibly owing to multiple instances of the same object.

object is moving, outputting a spatio-temporal tube that captures this most salient moving object. This paper, however, does not consider context within video for detecting multiple objects.

A few recent papers [1, 17] identify the important role of context in visual recognition. For object detection in images, Bell *et al*. [1] use spatial RNNs to harness contextual information, showing large improvements on PASCAL VOC [5] and Microsoft COCO [15] object detection datasets. Their approach adopts proposal generation followed by classification framework. This paper exploits spatial, but not temporal context.

Recently, Kang *et al*. [13] introduced tubelets with convolutional neural networks (T-CNN) for detecting objects in video. T-CNN uses spatio-temporal tubelet proposal generation followed by the classification and re-scoring, incorporating temporal and contextual information from tubelets obtained in videos. T-CNN won the recently introduced ImageNet object-detection-from-video (VID) task with provided densely annotated video clips. Although the method is effective for densely annotated training data, it's behavior for sparsely labeled data is not evaluated.

By modeling video as a time series, especially via GRU [2] or LSTM RNNs[11], several papers demonstrate improvement on visual tasks including video classification [28], activity recognition [4], and human dynamics [7]. These models generally aggregate CNN features over tens of seconds, which forms the input to an RNN. They perform well for global description tasks such as classification [4, 28] but require large annotated datasets. Yet, detecting multiple generic objects by explicitly modeling video as an ordered sequence remains less explored.

Our work differs from the prior art in a few distinct ways. First, this work is the first, to our knowledge, to demonstrate the capacity of RNNs to improve localized object detection in videos. The approach may also be the first to refine the object predictions of frame-level models. Notably, our model produces significant improvements even on a small dataset with sparse annotations.

5 Conclusion

We introduce a framework for refining object detection in video. Our approach extracts contextual information from neighboring frames, generating predictions with state of the art accuracy that are also temporally consistent. Importantly, our model benefits from context frames even when they lack ground truth annotations.

For the recurrent model, we demonstrate an efficient and effective training strategy that simultaneously employs localization-level strong supervision, category-level weak-supervision, and a penalty encouraging smoothness of predictions across adjacent frames. On a video dataset with sparse object-level annotation, our framework proves effective, as validated by extensive experiments. A subjective analysis of failure cases suggests that the current approach may struggle most on cases when multiple rapidly moving objects are in close proximity. Likely, the sequential smoothness penalty is not optimal for such complex dynamics.

Our results point to several promising directions for future work. First, recent state of the art results for video classification show that longer sequences help in global inference. However, the use of longer sequences for localization remains unexplored. We also plan to explore methods to better model local motion information with the goal of improving localization of multiple objects in close proximity. Another promising direction, we would like to experiment with loss functions to incorporate specialized handling of classification and localization objectives.

References

[1] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross B. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *to appear in CVPR 2016*, abs/1512.04143, 2015. URL http://arxiv.org/abs/1512.04143.

[2] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.

[3] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, pages 3061–3069, 2015.

[4] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014. URL http://arxiv.org/abs/1411.4389.

[5] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88 (2):303–338, June 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0275-4. URL http://dx.doi.org/10.1007/s11263-009-0275-4.

[6] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[7] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[8] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2014.

[10] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. URL http://arxiv.org/abs/1504.08083.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[12] Vicky Kalogeiton, Vittorio Ferrari, and Cordelia Schmid. Analysing domain shift factors between videos and images for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[13] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. *to appear in CVPR*, 2016. URL http://www.ee.cuhk.edu.hk/~wlouyang/Papers/KangVideoDet_CVPR16.pdf.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

[16] Zachary Chase Lipton, David C. Kale, Charles Elkan, and Randall Wetzell. Learning to diagnose with LSTM recurrent neural networks. *ICLR 2016*, 2016. URL http://arxiv.org/abs/1511.03677.

[17] Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Chen-Change Loy, and Xiaoou Tang. Deepid-net: Deformable deep convolutional neural networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[18] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3282–3289, 2012.

[19] Alessandro Prest, Vicky Kalogeiton, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Youtube-objects dataset v2.0, 2014. URL calvin.inf.ed.ac.uk/datasets/youtube-objects-dataset. University of Edinburgh (CALVIN), INRIA Grenoble (LEAR), ETH Zurich (CALVIN).

[20] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *to appear in CVPR 2016*, abs/1506.02640, 2015. URL http://arxiv.org/abs/1506.02640.

[21] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[22] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. URL http://arxiv.org/abs/1312.6229.

[23] Christian Szegedy, Scott E. Reed, Dumitru Erhan, and Dragomir Anguelov. Scalable, high-quality object detection. *CoRR*, abs/1412.1441, 2014. URL http://arxiv.org/abs/1412.1441.

[24] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL http://arxiv.org/abs/1605.02688.

[25] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, 2012.

[26] Subarna Tripathi, Serge J. Belongie, Youngbae Hwang, and Truong Q. Nguyen. Detecting temporally consistent objects in videos through object class label propagation. *WACV*, 2016.

[27] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li. Craft objects from images. *to appear in CVPR*, 2016. URL http://arxiv.org/pdf/1604.03239v1.pdf.

[28] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.