Supplementary Material of Efficient One-stage Video Object Detection by Exploiting Temporal Consistency

Guanxiong Sun[1,2], Yang Hua[1], Guosheng Hu[2], and Neil Robertson[1]

[1] EEECS/ECIT, Queen's University Belfast, UK
[2] Oosto, Belfast, UK
{gsun02,y.hua,n.robertson}@qub.ac.uk, huguosheng100@gmail.com

1 The Effect of SPN in Speed

One-stage detectors spend most time in decoding detections from every location on the low-level feature maps. We introduce the object size prior knowledge (SPN) to reduce the unnecessary computational cost wasted on the low-level feature maps in some video frames. The pipeline of SPN is described as follows: (1) At a time step t, SPN selects a set of reasonable bounding boxes $\{D_t^r\}$, whose classification score > 0.1, from all detected boxes $\{D_t\}$; (2) SPN records the level information for each bounding boxes in $\{D_t^r\}$, i.e., which feature level the bounding box is generated from, denoted as $\{L_t^r\}$; (3) For the next T time steps, SPN controls the detection heads to be conducted only on the levels in $\{L_t\}$. In summary, the speed improvement bring by SPN is mainly because of the skipping of computations on low-level feature maps.

Table S1. Distribution of the reasonable bounding boxes on different feature levels.

Level	#Box	%
1	44195	15.7
2	62942	22.3
3	97698	34.7
4	69215	24.6
5	7809	2.8

1.1 The Distribution of Objects On Different Feature Levels

Here we show the distribution of objects on different Feature Levels. Specifically, we use the pretrained FCOS [2] detector to predict bounding boxes on the ImageNet VID [1] validation set and accumulate the number of bounding boxes in $\{D_t^r\}$ on different feature levels for all frames. As shown in Table S1, only 15.7% of reasonable boxes are generated from feature level 1. In other words, SPN can address the computational bottleneck on feature level 1 for 84.3% frames.

1.2 Impact of SPN on Different Datasets

The impact of SPN for speed improvement can be different on other datasets, because of the different distribution of objects on low-level feature maps. For example, in the ImageNet VID dataset, most frames do not contain small objects and thus merely 15.7% boxes are generated from the level 1. In this case, SPN works well since the computation on level 1 for most frames can be skipped. However, for a dataset that dominated by small objects and most bounding boxes are generated from level 1, SPN may not improve the speed notably.

1.3 Qualitative Results of SPN

We showcase some detection results of FCOS [2] with and without using our SPN. As shown in Figure S1, the first column (a) shows the results of FCOS and the second column (b) shows the corresponding results of FCOS+SPN. In the first row where FCOS works well for the number 000012 frame. In this frame the best bounding box result comes from the third feature level as shown in the top left corner of the bounding box. However, in the following frames from 000013 to 000017, FCOS detects many false positive small bounding boxes from level 0. At the same time, the computations on the level 0 feature map are very heavy. In contrast, using FCOS+SPN, our method only detect object in the level 3 feature map for the number 000013 to 000017 frame. As a result, we can see less small false positives. Moreover, the runtime speed is significantly improved by skipping the computations on low-level feature maps.

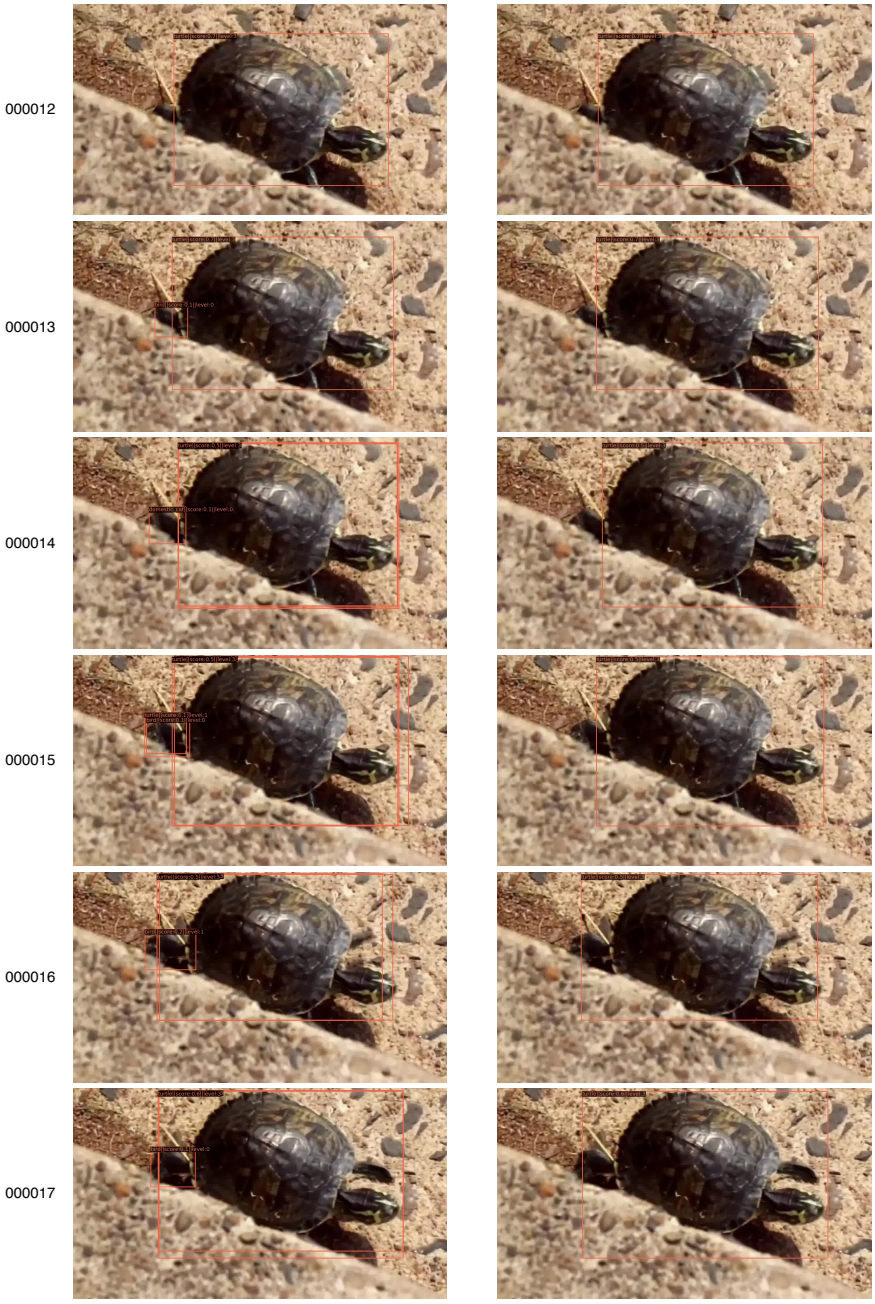

Fig. S1. Comparisons between detection results using FCOS and FCOS+SPN.

References

1. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
2. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: ICCV (2019)