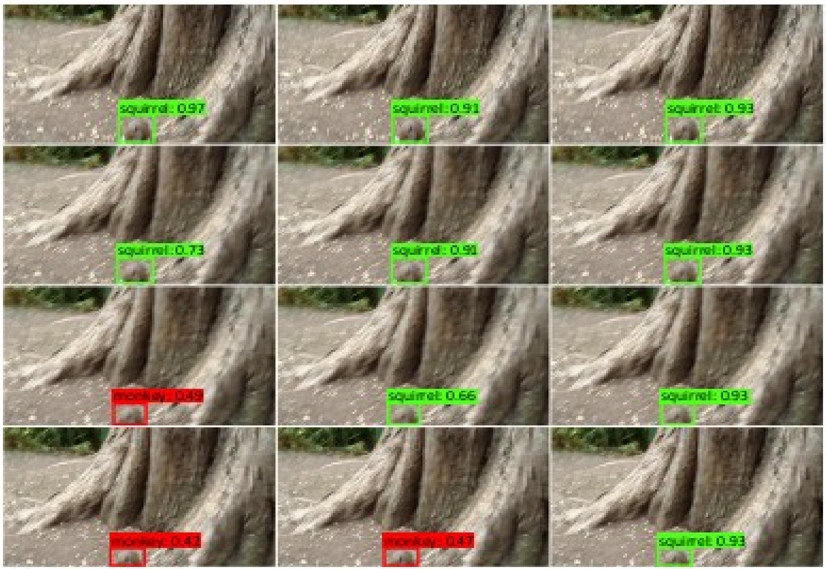
Object Detection in Videos by High Quality Object Linking

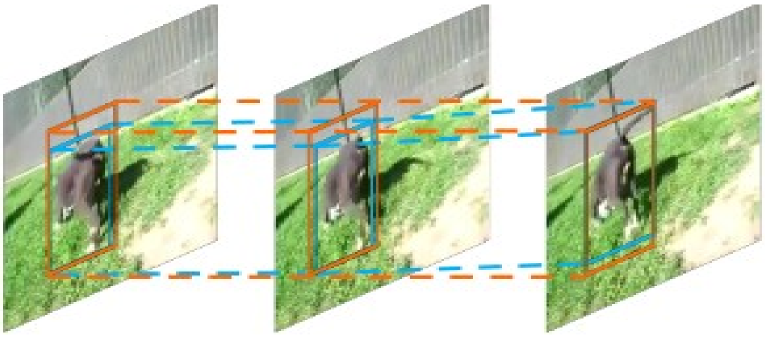
INTRODUCTION

Detecting of images in videos is a challenge due to degraded image qualities, e.g. motion blur and video defocus, leading to unstable classifications for the same object. Some of the methods like first use static image detectors to detect objects in each frame, and then link these detected objects by checking object boxes between neighboring frames, according to the spatial overlap between object boxes in different frames or predicting object movements between neighboring frames.



(a) (b) (c)

M ETHOD OR ALGORITHM



The task of video object detection is to infer the locations and classes of the objects in each frame of a video {I 1 , I 2 , . . . , I N } . To obtain high quality object linking, the method proposes to link objects in the same frame, which can be used to improve the classification accuracy.
Given a video divided into a series of temporally overlapping short video segments as the input, the method consists of three stages:
(1) Cuboid proposal generation for a short video segment.
(2) Short tubelet detection for a short video segment.
(3) Short tubelet linking for the whole video.

Cuboid Proposal Generation

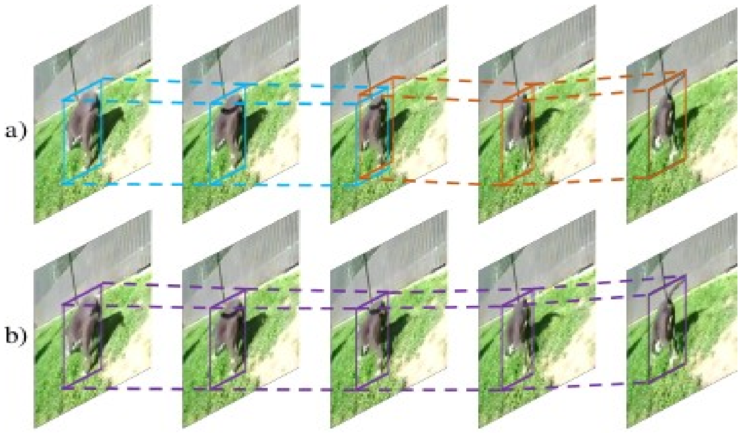
The region proposal network (RPN) method is modified in Faster R-CNN and introduce the cuboid proposal network (CPN) method for computing cuboid proposals. The output is a set of *whk* cuboid proposals, regressed from a $w \times h$ spatial grid, where there are k reference boxes at each location, and each cuboid proposal is associated with an objectness score.

Short Tubelet Detection

Used the 2D form of the cuboid proposal. Considering a frame Iτ in this segment,followed Fast R-CNN to refine the box and compute the classification score. Started with a RoI pooling operation, where the input is a 2D region proposal b and the response map of Iτ obtained through a CNN. The RoI pooling result is fed into a classification layer, outputting a {C + 1}-dimensional classification score vector yτ , where C is the number of categories and 1 corresponds to the background, as well as a regression layer, from which the refined box is obtained.

Short Tubelet Linking

The method divides a video into a series of temporally-overlapping short video segments of length K with stride K − 1 and performs a greedy short tubelet linking algorithm. Initially, put the short tubelets from all short video segments into a pool and record the corresponding segment for each tubelet. Our algorithm pops out the short tubelet T with the highest classification score from the pool. Check the IoU of the boxes if the IoU is larger than a threshold, fixed as 0.4, merge the two short tubelets into a single longer tubelet, remove the box with the lower score for the overlapping frame, update the classification score for the merged tubelet.



ADVANTAGES

Advantage in the linking in the same frame scheme : The approach is able to link the detected boxes in the same frame. The advantage in the linking in the same frame scheme is that there is no need to care about the object movement.

DISADVANTAGES

Linking the detected boxes obtained from static detector in neighboring frames might suffer from the object movement and harms the localization accuracy.

The two main issues that the approach has:

The multi-object issue : The potential way for the first issue is to generate multiple detection boxes from one proposal.

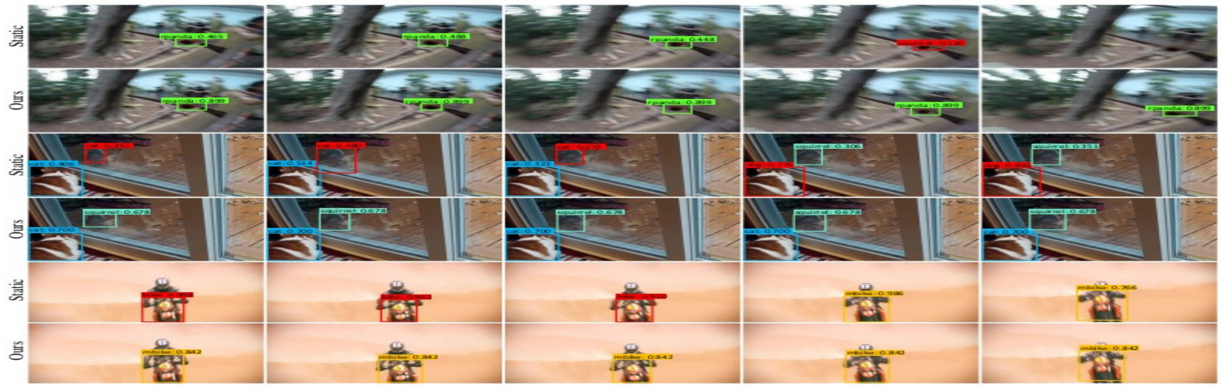
The boundary issue : The potential way for the second issue is to recheck the boxes in boundary frames separately.

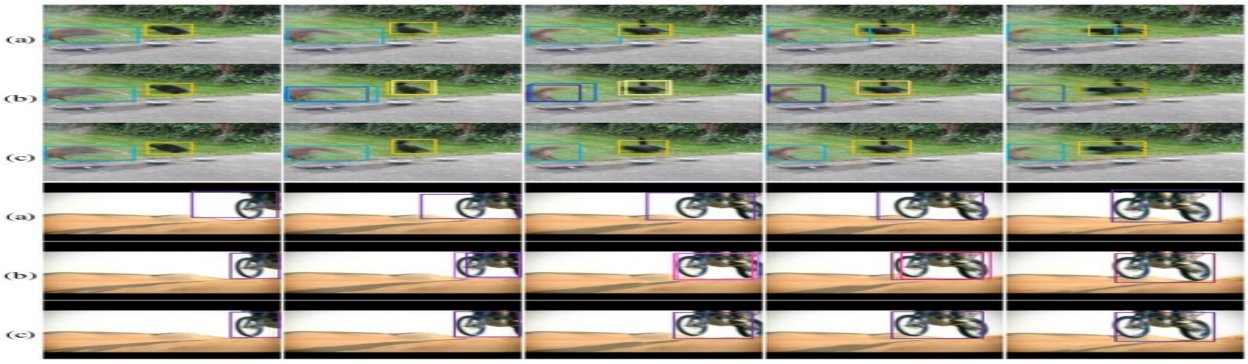
QUALITATIVE RESULTS

It can be seen that the static method fails to detect the red-panda when there are severe motion blurs and occlusions. This is reasonable because the appea- rance features have been severely degraded in this situation. After applying the object linking and rescoring, the method successfully classifies the target in the challenging frames.

Moreover, it is common that the static detectors may confuse with similar classes especially when a frame has low image quality and this can also be alleviated by rescoring the detections in the whole video because some frames have correct classifications and can propagate these scores to the challenging frames by object linking. To show that CPN enabling linking in the same frame leads to better localization accuracy. The fig below visualizes several object linking result comparisons between the method and the baseline approach of static detector + linking in neighboring frames over two examples. For (a), generated per-frame detection results using static image detector and then link detection boxes in neighboring frames by Seq-NMS. For (b), the results are from the approach without later short tubelet linking. For (c), the final results are from the approach with short tubelet linking.

Consequently, the approach is able to link the detected boxes in the same frame. The advantage in the linking in the same frame scheme is that there is no need to care about the object movement. In contrast, linking the detected boxes obtained from static detector in neighboring frames might suffer from the object movement and harms the localization accuracy.





CONCLUSION

Linked objects in the same frame for high quality object linking to improve the classification quality.

The method has three main components to achieve the goal:

(1) cuboid proposal network, (2) short tubelet detection, and (3) short tubelet linking.

The two main issues that the approach has: the multi-object issue and the boundary issue.

The potential way for the first issue is to generate multiple detection boxes from one proposal.

The potential way for the second issue is to recheck the boxes in boundary frames separately.