Deep Reinforcement Learning of Region Proposal Networks for Object Detection Supplementary Material

Aleksis Pirinen[1] and Cristian Sminchisescu[1,2]
[1]Department of Mathematics, Faculty of Engineering, Lund University
[2]Institute of Mathematics of the Romanian Academy
{aleksis.pirinen, cristian.sminchisescu}@math.lth.se

In this supplementary material we provide additional visualizations of how the drl-RPN agent operates on input images, see Fig. 1 - 3. Attached is also video visualizations of search strategies, showing search trajectories step-by-step.[1]

For the video visualizations, we show for each input image the search trajectories corresponding to three different exploration penalties[2] β. Specifically, we use $\beta = 0.025$, $\beta = 0.075$ and $\beta = 0.225$ (visualizations shown in this order). We provide visualizations on six input images from the PASCAL VOC 2012 test set, and we now briefly explain the respective visualizations.

Image 1: The image shows a part of a motorcycle. For small β, several fixations are performed (detecting the motorcycle at several fixation locations, and this context is accumulated by the agent for subsequent decision making and potentially for boosting final detection accuracy). The length of the search trajectory decreases with increasing β, and at $\beta = 0.225$ the agent is done already after the first fixation, confidently detecting the motorcycle.

Image 2: The image shows five sail boats. The length of the search trajectory decreases with increasing β. For the lower β, the objects may be locally detected at various fixation locations, and such class-specific context is used to guide the subsequent search process, and potentially to boost final detection accuracy. For $\beta = 0.025, 0.075$, all five boats are detected, although the left-most sail boat is detected twice (the boat without its sail is incorrectly detected). For $\beta = 0.225$, search stops already after two fixations, having successfully detected four out of the five boats (the small boat in the middle is missed).

Image 3: The image shows a clearly visible aeroplane. Independently of β, the aeroplane is confidently detected already after the first fixation. Thus, β only softly enforces the exploration extent – in very simple images the agent may still terminate the search early despite being given a small β. In general, given β, the search remains image- and category-dependent.

Image 4: The image shows a person and a dog. The length of the search trajectory decreases with increasing β, and at $\beta = 0.225$ both objects are confidently detected after two fixations. For the lower β, the objects may be locally detected at various fixation locations, and such class-specific context is used to guide the subsequent search process, and potentially to boost final detection accuracy.

Image 5: The image shows two aeroplanes and a person. The length of the search trajectory decreases with increasing β, and at $\beta = 0.225$ all three objects are confidently detected after two fixations. For the lower β, the objects may be locally detected at various fixation locations, and such class-specific context is used to guide the subsequent search process, and potentially to boost final detection accuracy.

Image 6: The image shows four cars (in addition to some barely visible cars in a traffic jam), mainly viewed from the back. The length of the search trajectory decreases with increasing β, and at $\beta = 0.225$ all four cars are confidently detected after three fixations. For the lower β, the cars may be locally detected at various fixation locations, and such class-specific context is used to guide the subsequent search process, and potentially to boost final detection accuracy.

[1]Note that the visualizations do *not* reflect the runtimes of our sequential detector; the updates are shown slowly to make it easier to follow.

[2]Recall from the main paper that the models trained with adaptive exploration penalties do *not* use posterior class probability adjustments. Such adjustments could however be applied within such models as well, which is why they are mentioned in the video visualizations.

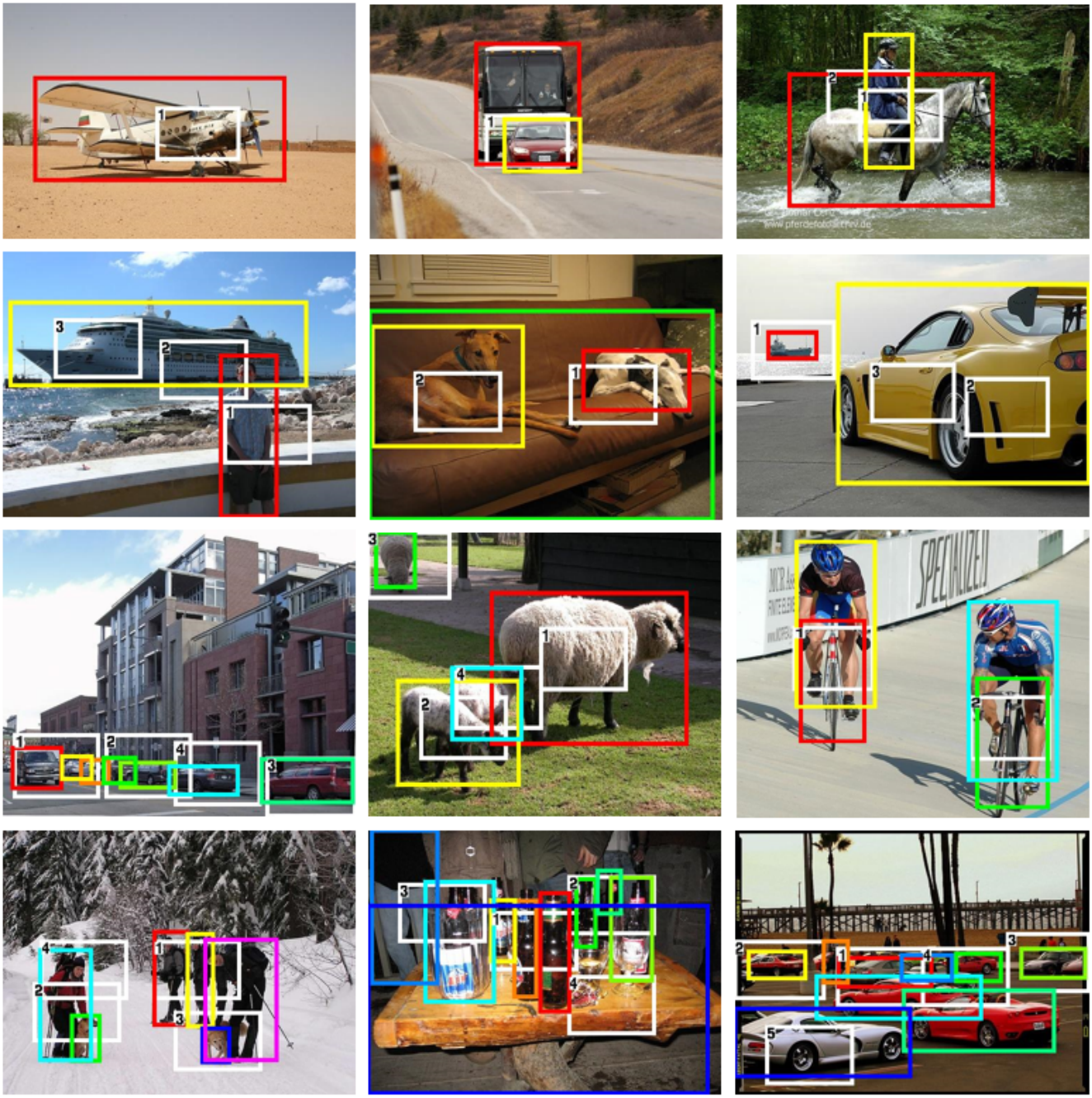

Figure 1: Upscaled fixation areas (*attention boxes*, white) generated by our sequential search model drl-RPN, together with final detection boxes (colored), for several images from the PASCAL VOC 2007 test. Each attention box has an associated number, showing at which time-step $t \geq 1$ the corresponding area in the feature map was observed. Depending on the complexity of a scene, such as the number and layout of objects, the model automatically determines when to stop the search process. The top two rows show examples of short search trajectories, in which only one or a few object instances exist in the image. In contrast, longer trajectories are shown in the bottom two rows, corresponding to images containing more object instances and/or categories. As such, the number of *fixate* actions is *not* necessarily equal to the number of object instances, but depends also on the layout of the objects (*e.g.*, how close objects are to each other). For example, in the top-mid image, only one *fixate* action is necessary to simultaneously locate the bus and the car, whereas an additional fixation is produced for the image to the right on the second row. Overall however, the number of fixations typically increases with the number of object instances, as would be expected. Note that the fix-sized attention boxes are *not* in any way related to the sizes of the RoIs being forwarded for class-specific predictions. These boxes only correspond to what subset (and where) of RoIs are selected.

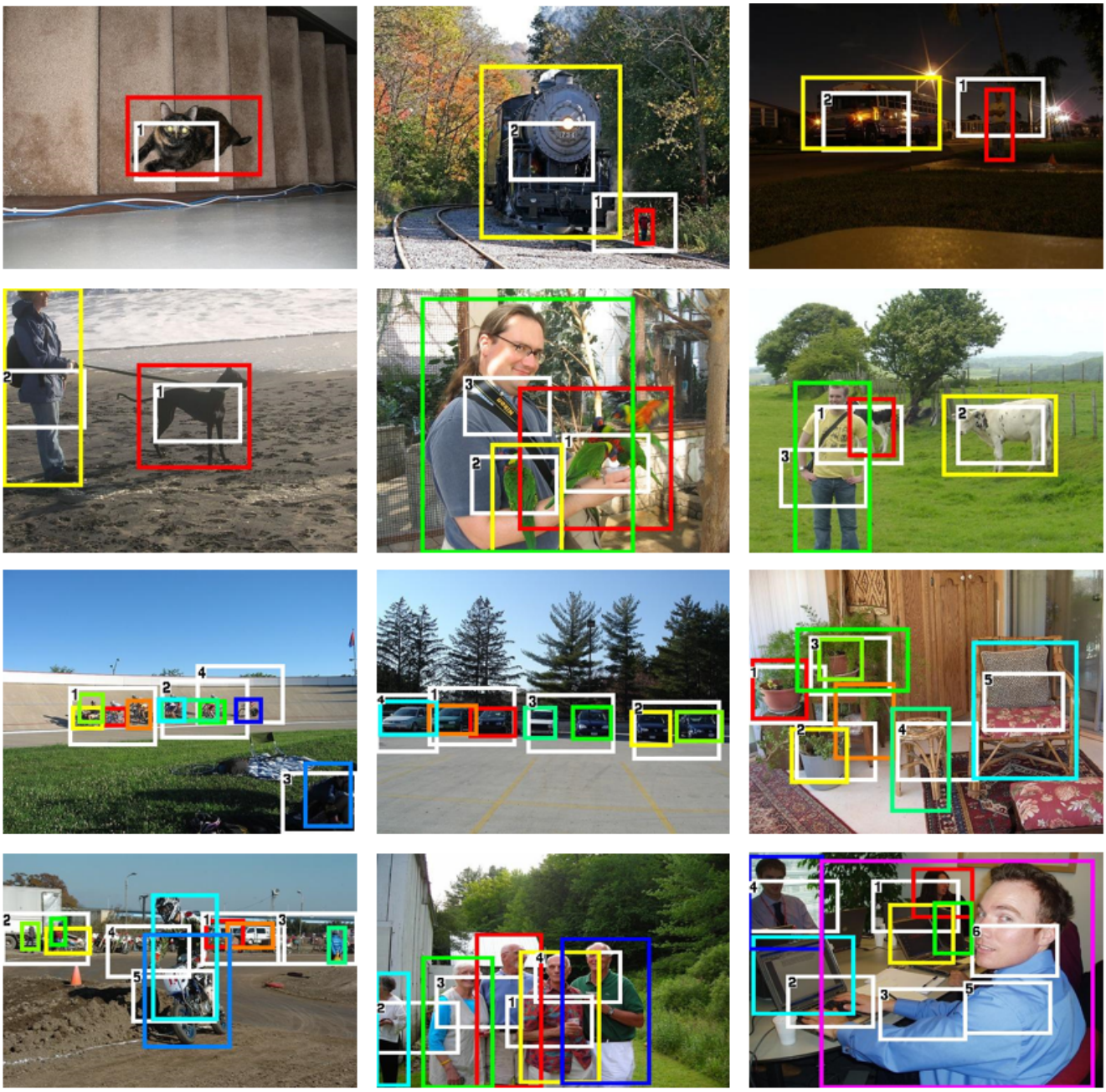

Figure 2: Additional visualizations of successful search strategies of our drl-RPN model on the PASCAL VOC 2007 test, c.f. Fig. 1. Depending on the complexity of a scene, such as the number and layout of objects, the model automatically determines when to stop the search process. The top two rows show examples of short search trajectories, in which only one or a few object instances exist in the image. In contrast, longer trajectories are shown in the bottom two rows, corresponding to images containing more object instances and/or categories. Note how our model is able to adapt its strategy to a variety of situations. For example, in the top-mid image the large train and tiny person are both elegantly captured, and in the poorly illuminated image to the top-right both the bus and person are discovered in what appears to be the smallest possible number of fixations. The distance between different fixation locations can vary quite drastically too. An example of this is seen in left image of the third row: the agent moves its view from the cluster of bicycles (fixations 1 - 2) to the barely visible person sitting in the shadow (fixation 3), and then back to investigating the bicycles (fixation 4).

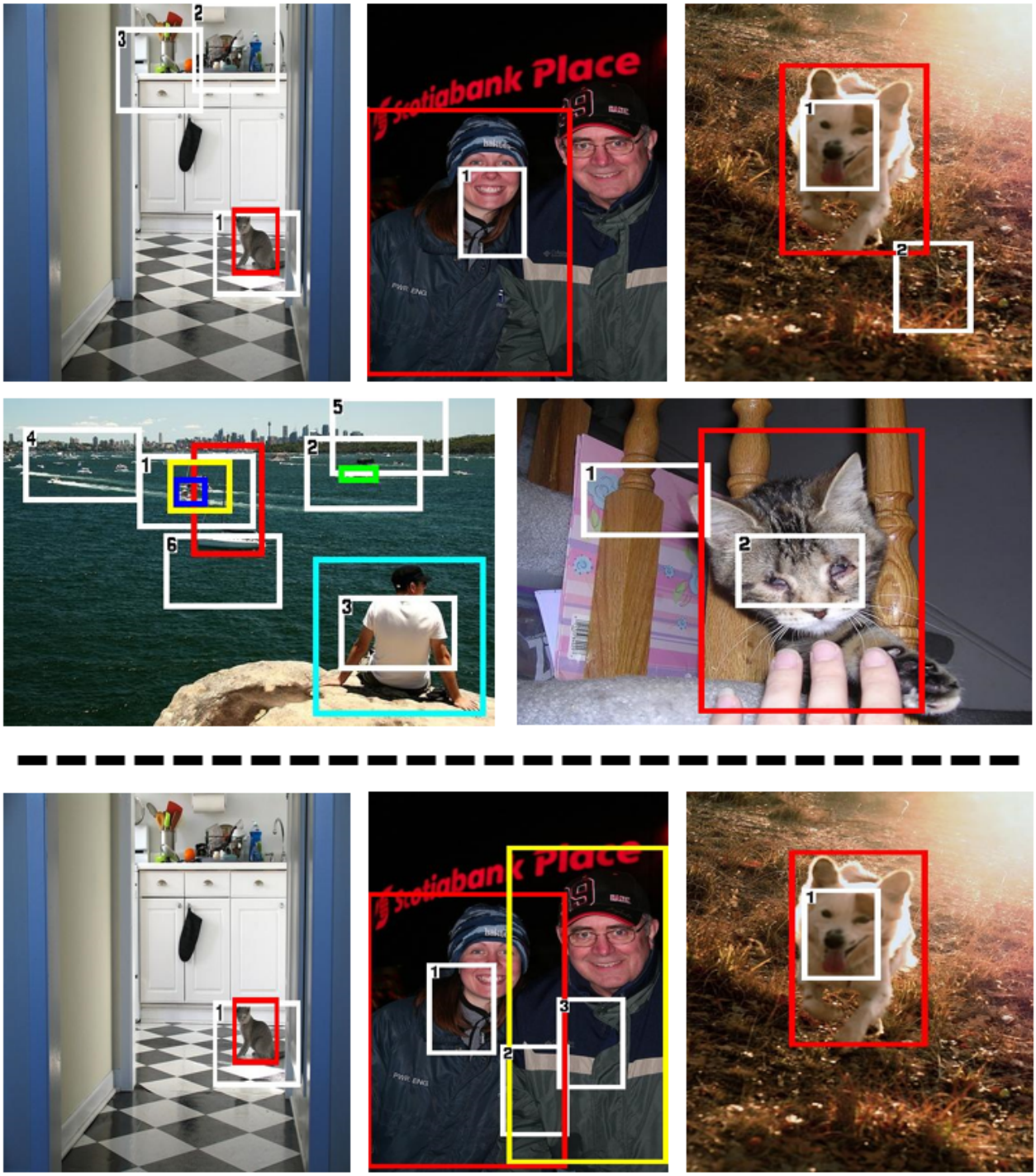

Figure 3: Similar visualizations of search strategies as in Fig. 1 - 2; in this case a few examples of slightly less successful and/or unexpected search trajectories are shown (above dashed line) on the PASCAL VOC 2007 test. The model sometimes appears to do one or a few additional, unnecessary fixations (such as in the image of the dog to the top-right, in which the dog is detected already at the first fixation). It may also be the case that the additional fixations occur at locations which are object-like in a more generic sense (”stuff”). An example of this can be seen in the image of the cat to the top-left, with two additional fixations at the kitchen counter, which contains several items that are not labeled in the training data. Similarly, on the mid-left the agent searches among all the tiny boats in the distance, of which only a few are detected in the end. Finally, the model may occasionally stop the search too early, as is apparent in the top-mid figure, in which the man to the right is not detected. Below the dashed line are the corresponding top images where the stopping condition has manually been altered to perform more/less fixations.