Attend and Segment: Supplementary Material

Soroush Seifi[0000−0002−4791−5350] and Tinne Tuytelaars[0000−0003−3307−9723]

KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{FirstName.LastName}@esat.kuleuven.be

1 Baselines

Figure 1 complements figure 9 from the main paper by demonstrating the results on the Camvid and Kitti datasets. According to this figure, the same arguments as discussed in section 4.2 of the main paper hold for these datasets too.

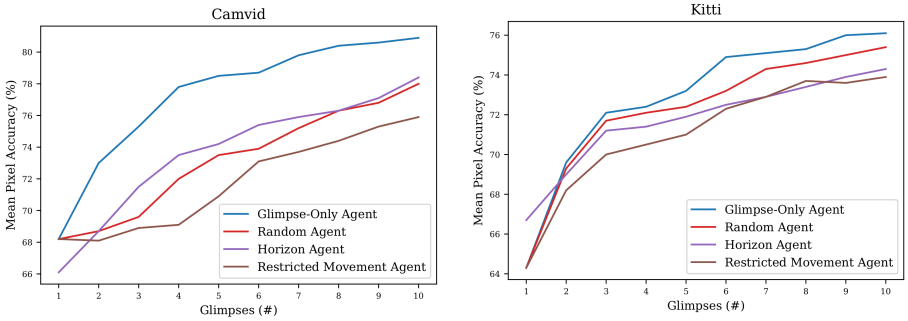

Fig. 1. Comparison with the baselines defined in section 4.2 on Camvid and Kitti Datasets.

2 Hybrid Agent 2

To further study the hybrid agent defined in section 4.3 (hybrid agent 1), we define another hybrid agent which starts from a downscaled image with a much lower resolution (hybrid agent 2). It can see the whole environment in 16×8 pixels which is 8 times smaller than the resolution for hybrid agent 1 (and 256 times smaller than the input image). This downscaled image costs 128 pixels of the agent's pixel budget which is smaller than a quarter of number of pixels in one 3-scales retina glimpse. Hybrid agent 1's downscaled view cost 1024 pixels equal to almost two retina glimpses.

As seen in figure 2, hybrid agents can improve the performance where the glimpses' coverage is low. However, as the number of glimpses and consequently coverage increases the gap between glimpse-only and hybrid agents gets smaller. On Cityscapes and Camvid datasets the glimpse-only agent even exceeds hybrid agent 2's performance after 7 glimpses. For experiments where the number of glimpses is low, the hybrid-agent can still rely on the 16×8 view of the whole input to predict the structure of the areas not seen in the glimpses. However, for the experiments with a high pixel budget, the agent can take glimpses of

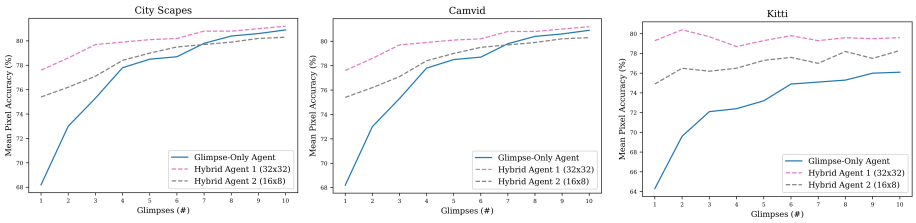

Fig. 2. Hybrid agents' performance comparison against the glimpse only agent. .

the most uncertain areas with higher resolution compared to the 16×8 view and therefore the glimpse-only agent may outperform hybrid agent 2 in those cases. This also explain why all agents converge to almost the same result after 10 glimpses.

Figure 1 and 2 suggest that the glimpse-only agent's performance is lower on Kitti compared to the other two datasets. The limited number of training examples (160 training and 40 test images in our experiments) prevents the glimpse only agent to generalize well without access to the full image. Therefore, even a very low resolution view of the environment can always boost its performance. Consequently, hybrid agents are also useful in scenarios where the number of training examples are limited.

3 Glimpse-only Agent's Analysis

Figure 3 provides more examples on the outputs of the glimpse-only agent's modules. Our agent produces its predictions for the whole environment even after visiting only the small area covered by the first glimpse. Similarities in the predictions after the first step can be a good indication of the prior knowledge learned by our agent about the average structure of the dataset (CitySapes in this case). It can be inferred from figure 3 that this dataset consists mostly of street images with queues of cars parked to the side of the street since this is similar for all predictions after the first step where agent's knowledge of the environment is very limited. Furthermore, the certainty maps generated after the first step indicate that with a high confidence there is a similar semi-circle like area in front of the car for all images in the dataset. This semi-circle consists of the street and the Mercedes-Benz sign. Therefore, a heuristic for saving a significant amount of computation is to avoid processing the pixels in this area. This is learned by our agent without a need for heuristics.

Another interesting property is that the agent chooses the next location purely based on the information it gained from the previous glimpses. Although the first glimpse in 3 examples shown in figure 3 are very close to each other, the next glimpse is sampled from totally different locations in each case. This means that our agent reacts differently to different environments based on the information it has received by the glimpses and not only by the location of them.

Finally, its worth noting that the final prediction at each step combines and refines local, global and the last step's predictions. It relies on the global module to fill-in the unvisited areas and the local module for the visited parts while lowering the noise of the segmentations produced by each one of these modules.

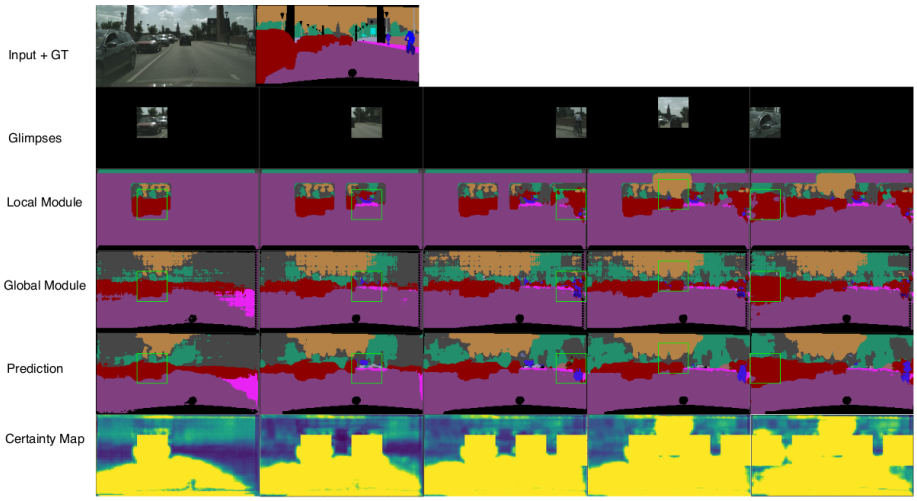

Fig. 3. Outputs of the glimpse-only agent's modules for 5 steps.