A Computer Vision-Based Attention Generator using DQN

Jordan Chipka
General Motors
jbc274@cornell.edu

Shuqing Zeng
General Motors
shuqing.zeng@gm.com

Thanura Elvitigala
General Motors
thanura.elvitigala@gm.com

Priyantha Mudalige
General Motors
priyantha.mudalige@gm.com

Abstract

A significant obstacle to achieving autonomous driving (AD) and advanced driver-assistance systems (ADAS) functionality in passenger vehicles is high-fidelity perception at a sufficiently low cost of computation and sensors. An area of research that aims to address this challenge takes inspiration from human foveal vision by using attention-based sensing. This work presents an end-to-end computer vision-based Deep Q-Network (DQN) technique that intelligently selects a priority region of an image to place greater attention to achieve better perception performance. This method is evaluated on the Berkeley Deep Drive (BDD) dataset. Results demonstrate that a substantial improvement in perception performance can be attained – compared to a baseline method – at a minimal cost in terms of time and processing.

1. Introduction

As AD/ADAS features continue to advance, vehicles are being equipped with an increasing number of sensors and the demand for high-definition (HD) sensing is likewise rising [6]. However, due to limited computing resources, this demand for HD sensing is often unworkable for real-time systems [16]. Image-based object detection has a time complexity of $O(N)$. Therefore, this results in a tradeoff being negotiated between perception performance and processing time/power. While there is currently a large emphasis on developing more powerful embedded computing platforms for AVs, an alternative approach of addressing this difficulty is to develop more intelligent methods of processing sensor data.

A common technique is to mimic human foveal vision by only processing select parts of the scene in HD when and where it is needed, rather than continually processing the entire scene [10, 1]. The most common method of multi-resolution sensing employed in the automotive industry is to use two front-facing cameras in tandem – one wide field-of-view (FOV) low-resolution camera for close-range object detection and one narrow FOV high-resolution camera for long-range object detection. A less expensive attention tactic is to use one camera but process the full camera frame at a downscaled resolution and process a small center crop of the frame at a higher resolution. This approach demands a smaller pixel budget while allowing for HD detections in the center of the frame to be obtained. However, these approaches have disadvantages in that they do not ensure that the HD imagery is being taken from the most relevant region of the image (e.g., when on curved roads, hills, intersections, etc.).

More sophisticated attention techniques typically fall into two categories: bottom-up and top-down attention models. A bottom-up attention model uses features or characteristics from the scene to derive its attention. A classic illustration of bottom-up attention involves a human subject having their attention drawn to a single horizontal bar in a scene filled with vertical bars [25]. Top-down attention models, on the other hand, are based on prior knowledge and are based on goals, expectations, and/or rewards. A well-known example of top-down attention involves human subjects watching a scene of a family when an unexpected visitor enters into a room. Before watching the scene, the subjects are given different tasks such as predicting the material circumstances of the family, estimating the ages of the people, and freely examining the scene. Based on their given tasks, the human subjects demonstrate considerably different eye movements [30].

Quickly and automatically detecting relevant areas of a scene through bottom-up attention models is an appealing capability for machine vision [19]. Numerous studies in recent years have used bottom-up attention models for the tasks of object segmentation, object recognition, image captioning, and visual question answering [8, 29, 24, 7]. These

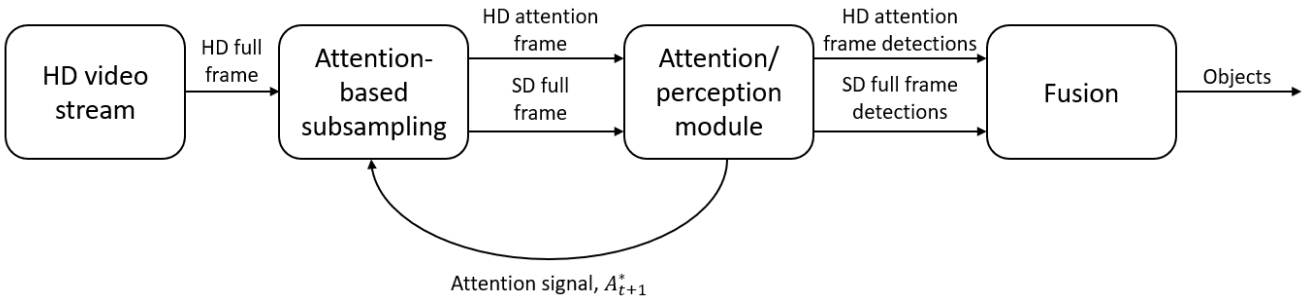

Figure 1. Workflow of the proposed attention-based sensing scheme.

methods typically use convolutional neural networks and/or recurrent neural networks to identify the attention region – which is not favorable for safety-critical systems with real-time constraints due to additional computing needs. Another common bottom-up approach is to identify salient regions in the image [9, 2, 11]. However, for the application of autonomous driving, salient regions are not necessarily the most relevant regions of the scene. Salient regions often identify the most conspicuous objects in the scene, however these objects do not require HD sensing as they could also be detected with low resolution sensing. Rather, relevant regions in a roadway scene can include small distant objects such as an upcoming traffic light or pedestrians crossing at an upcoming crosswalk.

Although the majority of attention models use a bottom-up approach, it is widely accepted that top-down factors play a key role in attention guidance [12]. A recurrent attention model (RAM) and deep recurrent attention model (DRAM) were proposed to mimic human attention and have demonstrated promising results for the tasks of image classification and digit recognition [20, 3]. Additional saliency-based techniques have also taken a top-down approach to deriving attention and predicting human gaze [13, 14, 28]. Finally, spatial transformers have been used as an attention mechanism for digit classification tasks due to their advantage of being fully differentiable [15, 18].

Taking inspiration from recent advances in computer vision-based reinforcement learning (RL), in particular, Deep Q-Network (DQN) [17], as well as aspects from both bottom-up and top-down attention models, we present a lightweight DQN agent that intelligently selects the most relevant region of the image to give more attention and demonstrate its benefit on the task of object detection on roadway scenes. This approach is a hybrid between bottom-up and top-down attention because it uses low-level visual features as its input, but is trained to achieve a high-level goal through the DQN agent's reward function. This DQN agent emulates human foveal vision by using low-resolution for peripheries but high-resolution for critical areas, and therefore reduces the overall number of pixels to process. This approach offers a more optimized technique to address the tradeoff between efficiency and effectiveness. Finally, this method is self-contained, derived only from perception, and therefore can be easily integrated into any existing perception system.

2. Methodology

This work aims to develop an intelligent attention-subsampling technique in which the required compute resources are significantly reduced while the perception performance is retained. Figure 1 shows the workflow for the proposed technique in which human foveal vision is emulated as a bi-directional process. In this workflow, an HD image is captured from the camera stream and passed on to the attention-based subsampling mechanism. The subsampling mechanism first downscales the full camera frame to a standard-definition (SD) resolution and then passes it along to the DQN-based attention module. This module then identifies the most relevant region of the image and passes that back to the attention-based subsampling module in the form of an attention signal. Based on the attention signal, the subsampling mechanism crops out an attention window from the original HD frame and passes that on to the perception module. Finally, object detection is performed on both the SD full-frame and the HD attention window, and the resulting mixed-resolution detections are fused to yield the final objects.

2.1. DQN-Based Attention Agent

The first step to developing the DQN-based attention agent is establishing its action space. The action, A^*_t, of the attention agent is simply the selected location of the attention window's centroid (i.e., x and y locations in image coordinates). To simplify the action space, we discretize the image into $N_x \times N_y$ blocks. Then, a $R_{N_x} \times R_{N_y}$ attention window is determined based on the given action, A^*_t. Figure 2 depicts the discretized image and the attention agent's action space.

The attention/perception module shown in Figure 1

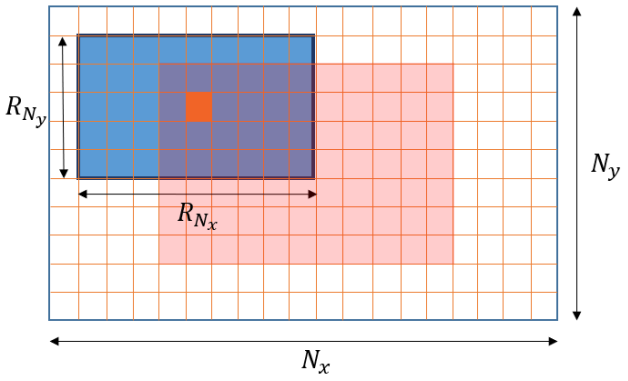

Figure 2. Depiction of the discretized image, attention region (blue), centroid of attention region (orange), and action space (red). The action space does not encompass the entire image in order to ensure the attention region never extends beyond the image boundary.

is mainly comprised of a CNN-based object detection (OD) network, as well as a lightweight deep Q-network (DQN) [4]. Figure 3 shows a detailed view of the attention/perception module and how it fits into the overall workflow. Object detection is performed twice sequentially – once for the SD frame and then once for the HD attention frame. First, the SD full-frame is passed into the OD network and its detections are then received out from the OD head. YOLOv5, the latest version in the YOLO object detector family, was used in this study [21, 22, 23, 5, 26]. Before being passed to the OD head, the resulting feature tensor from the CNN backbone is intercepted to be used as the input to the DQN. The output of the DQN is an array of Q-values, which for our application, represents the anticipated reward for each possible location of the attention window, A_t. The argmax of this array is then taken to obtain the attention signal, A_t^*. Finally, based on this attention signal, an HD attention frame is cropped from the original HD image, passed to the OD network, and HD attention frame detections are then received and fused with the SD full-frame detections to yield the final objects.

The architecture of the DQN consists of two branches, both of which are fully connected neural networks comprised of one hidden layer. One branch is used to select the x location of the attention window, while the other branch is used to select the y location of the attention window. The architecture of the DQN is depicted in Figure 4. The input to the DQN is the feature tensor from the object detection backbone. This feature tensor is then flattened and passed to two distinct, fully connected networks. The first network outputs an array with size N_x. This output's maximum value represents the predicted column for the attention window's centroid that would maximize the reward. The second network outputs an array with size N_y. The maximum value of this output represents the predicted row for the attention window's centroid that would maximize the reward.

2.2. Training Process

This work uses standard techniques for training the DQN but with a few minor adjustments. The DQN is trained to approximate a function that can predict Q-values, which are a measure of the agent's expected reward, given an action A_t, state X_t, and network parameters θ. The policy which governs the agent's actions, $\pi_\theta(A_t, X_{t-1})$, is an epsilon greedy policy around the function $A_t^* = \arg\max_{A_t} Q(A_t)$ [27]. This policy is used to improve training stability and increase the likelihood of convergence.

The DQN was trained using the Berkeley Deep Drive (BDD) dataset [31]. The images contained in the BDD dataset are not sequential, and hence each image could be considered independent from the next. Therefore, a replay memory was not needed to randomly sample data in order to ensure decorrelated batches. Also, as a result of dealing with temporally independent data, the discount factor, γ, was set to be 0. The discount factor is a constant between 0 and 1 that determines how the agent is rewarded – i.e., a discount factor of 0 will reward the agent simply based on the instantaneous reward, while a value closer to 1 will give more significance to expected rewards in the future. Since our DQN essentially operates on a frame-by-frame basis, our agent is only concerned with maximizing the reward for the current frame, and therefore, the discount factor is set to 0.

2.2.1 Loss Function

With a discount factor of 0, the original training rule (as shown in Eq. 1) simplifies to Eq. 2.

$$Q^\pi(X, A) = r + \gamma Q^\pi(X', \pi(X')) \tag{1}$$

$$Q^\pi(X, A) = r \tag{2}$$

$$\delta = |Q^\pi(X, A) - r| \tag{3}$$

In the above equations, $Q^\pi(X, A)$ are the Q-values according to policy π, given state X and action A. The reward is represented as r, and the expected future state is X'. The difference between the two sides of the equality in Eq. 2 is the error, δ, that we are trying to minimize during training (as shown in Eq. 3). Since the DQN has two branches, one for the attention window's x position and the other for its y position, the network produces two sets of Q-values for

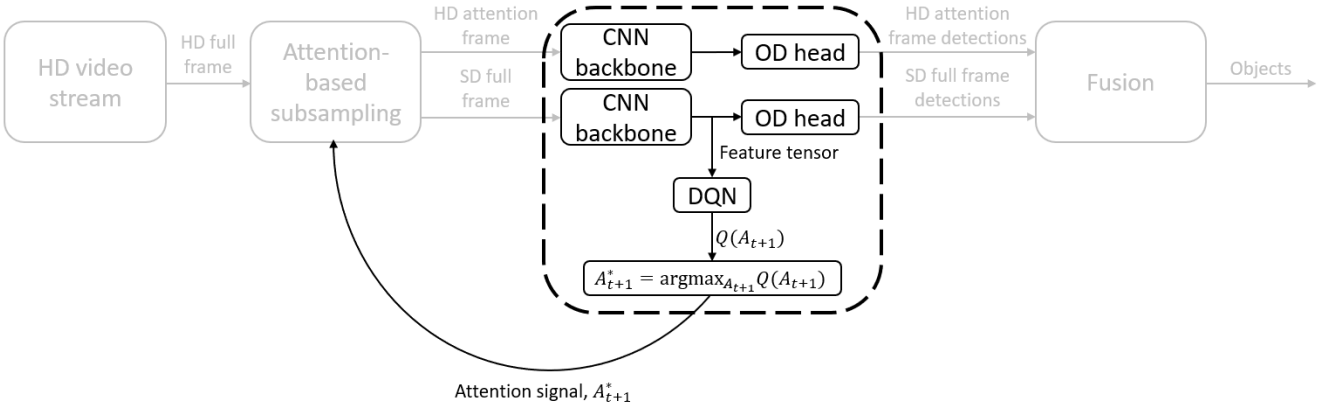

Figure 3. Detailed view of the attention/perception module and how it fits in the overall workflow.

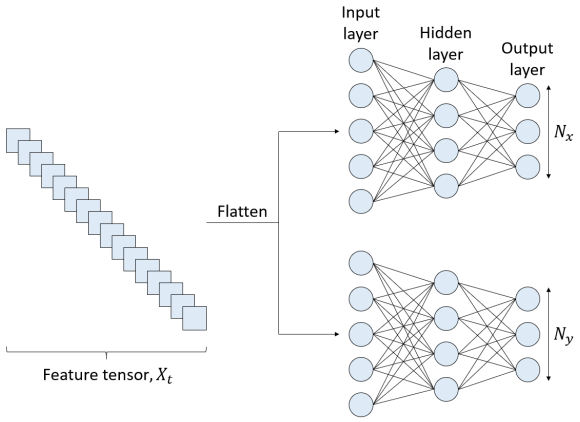

Figure 4. The network architecture of the DQN consists of two fully connected branches – one to predict the x location of the attention window and another to predict the y location of the attention window.

each batch during training. Therefore, we get two error values, δ_x and δ_y. Using an L1 loss function, we show the final training loss in Eq. 4, where B is the batch size.

$$\mathcal{L} = \sum_{i=0}^{B} (\delta_x + \delta_y) \tag{4}$$

2.2.2 Reward Function

A simple reward function was implemented to train the DQN agent. This reward function compared the number of true positive detections before applying the attention window, TP, to the number of true positive detections after applying the attention window, TP'. If applying the attention window resulted in more true positive detections, then the reward was +1. If applying the attention window resulted in the same amount of true positive detections, then the reward was 0. Finally, if applying the attention window resulted in less true positive detections, then the reward was -1. This reward function is summarized in Eq. 5.

$$r = \begin{cases} -1 & TP' - TP < 0 \\ 0 & TP' - TP = 0 \\ 1 & TP' - TP > 0 \end{cases} \tag{5}$$

For this study, the reward function was designed in a way to simply maximize true positive detections of any object. As will be shown in section 3 of this paper, the DQN agent learned to seek out small, distant vehicles using its attention window. This is due to vehicles being the dominant class in the BDD dataset. However, a more sophisticated reward function can easily be tailored in order to give more significance to other classes that may be of more interest. Furthermore, additional post-processing techniques could be implemented to give more significance to other regions based on other sources of information, such as a map, path planning, etc.

3. Results

3.1. Experimental Setup

The DQN-based attention mechanism was tested on both sequential and non-sequential data. For non-sequential data, only one camera frame could be used throughout the attention workflow. In other words, the attention window is applied to the same camera frame that provided the input feature tensor to its DQN. However, for sequential data, this process spans two frames so that the camera frame at time t is used to generate the attention window that is applied to the camera frame at time $t + 1$.

The experimental setup consists of comparing the performance of YOLOv5 at four different image resolutions. The first resolution is the original HD image resolution of 0.92 MP. Second, a downscaled SD image resolution of 0.25

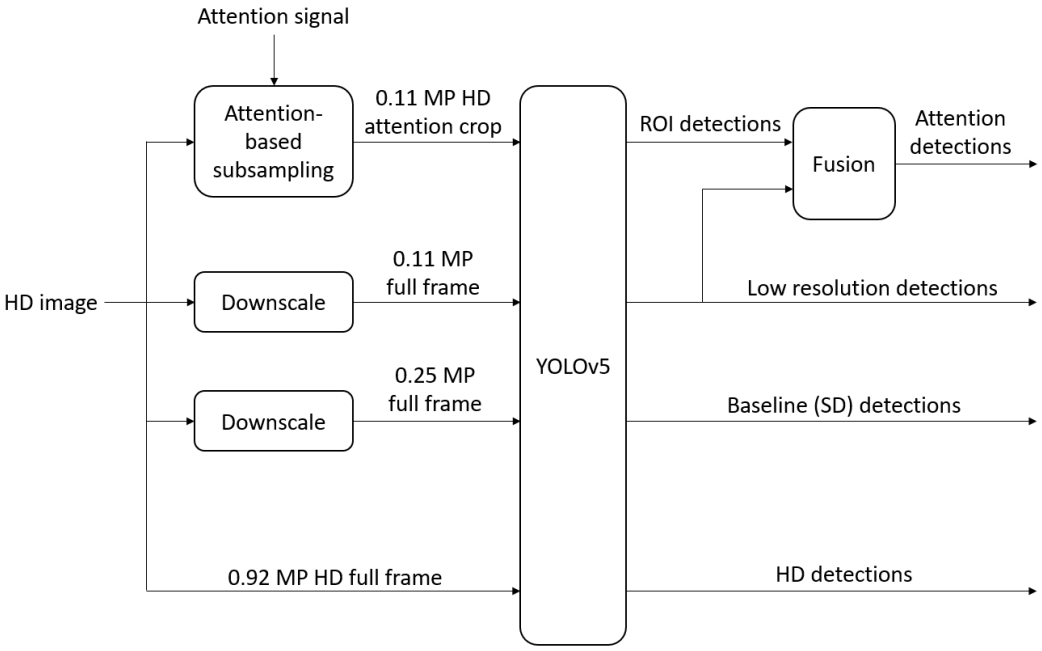

Figure 5. Summary of the experimental setup.

MP is used as the baseline. The third resolution is a further downscaled resolution of 0.11 MP. Finally, a 0.11 MP HD attention crop of the original image, referred to as the region of interest (ROI), is combined with the previously mentioned 0.11 MP downscaled frame to yield a 0.22 MP attention frame. A summary of this experimental setup is depicted in Figure 5.

3.2. Evaluation

Average precision (AP) was used as the performance metric for this study. Inference time was also used as a secondary performance metric. In addition to the DQN-based attention model, two additional attention models were used for comparison. First, a centered attention model was tested, which always placed the attention window in the middle of the camera frame. Second, a random attention model was tested, which randomly placed the attention window within the camera frame. Four classes were used during the evaluation – vehicle, traffic light, pedestrian, and bicycle. Finally, a weighted mean average precision (wMAP) was used to calculate the average precision over all classes, according to their prevalence in the dataset. Figures 6-10 show the wMAP and AP for various classes as a function of intersection over union (IOU). The inference time of the various methods is shown in Figure 11.

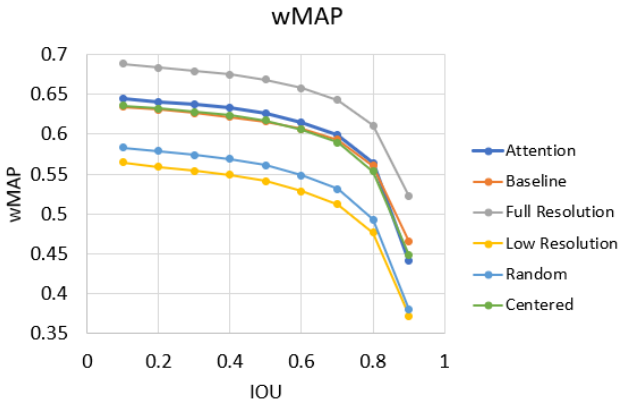

Figure 6. Weighted mean average precision as a function of IOU for the BDD dataset.

4. Discussion

4.1. Impact on Inference Time

The results in the previous section demonstrate that the proposed attention mechanism can provide a considerable performance boost in terms of average precision while only requiring a small increase in processing time, as shown in Figure 11. Although the total amount of pixels being processed for the attention method is fewer than for the baseline method (0.22 MP vs. 0.25 MP), the attention method requires YOLOv5 to be run twice to provide the detections for the low-resolution frame, as well as the region of in-

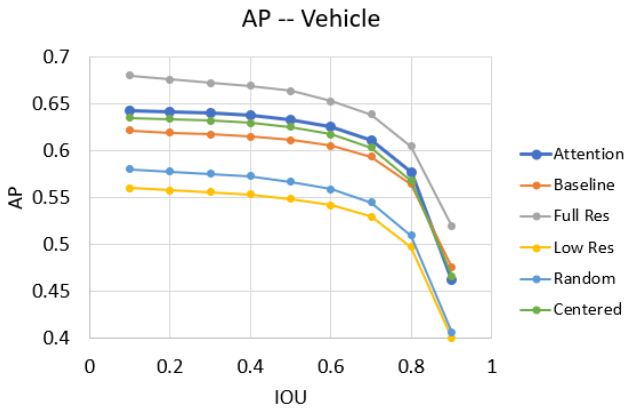

Figure 7. Average precision for vehicles as a function of IOU for the BDD dataset.

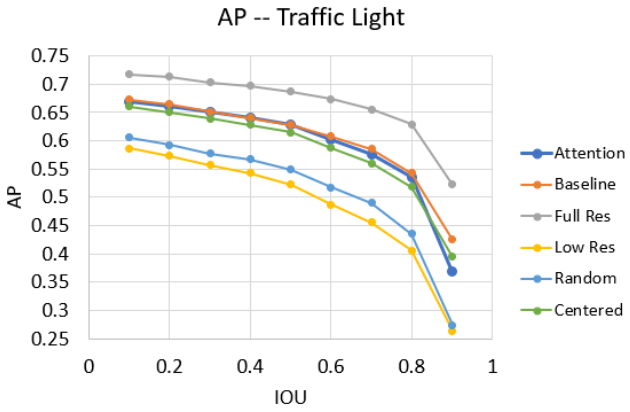

Figure 8. Average precision for traffic lights as a function of IOU for the BDD dataset.

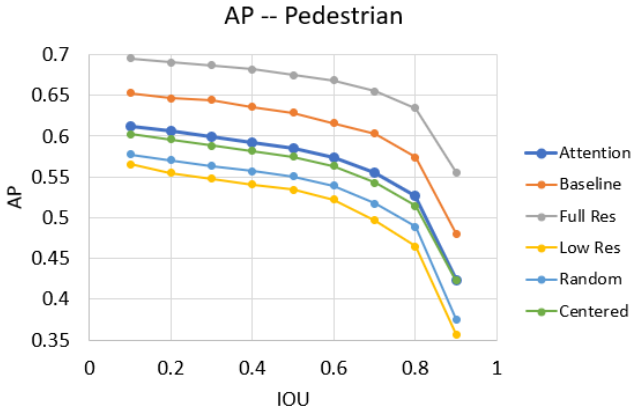

Figure 9. Average precision for pedestrians as a function of IOU for the BDD dataset.

terest (ROI) frame. Therefore, the extra overhead involved with running the object detection network twice results in

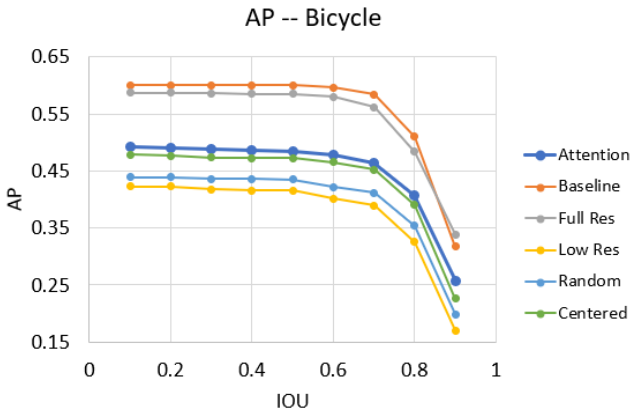

Figure 10. Average precision for bicycles as a function of IOU for the BDD dataset.

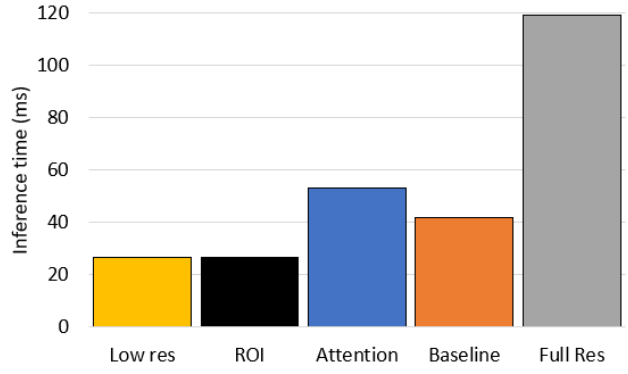

Figure 11. Inference time of YOLOv5 object detector at various resolution levels.

a slightly higher inference time; however, it is still much faster (less than half the required processing time) than the full resolution (HD) method. To further reduce processing time, a more sophisticated CNN architecture could be developed to process both frames simultaneously.

4.2. Impact on Detection Accuracy

As shown in Figure 7, the attention method provides a considerable improvement in performance over the baseline method for the vehicle class. The difference between the baseline results and the full resolution (HD) results – which represent the upper bound that can be achieved with the given data and object detection network – is nearly cut in half when the attention mechanism is used. However, the performance boost that the attention mechanism provides for wMAP is not as considerable. This is due to the lower performance of the attention method for the traffic light class – which performed comparably to the baseline method – and the pedestrian and bicycle classes – which both performed worse than the baseline method. Although

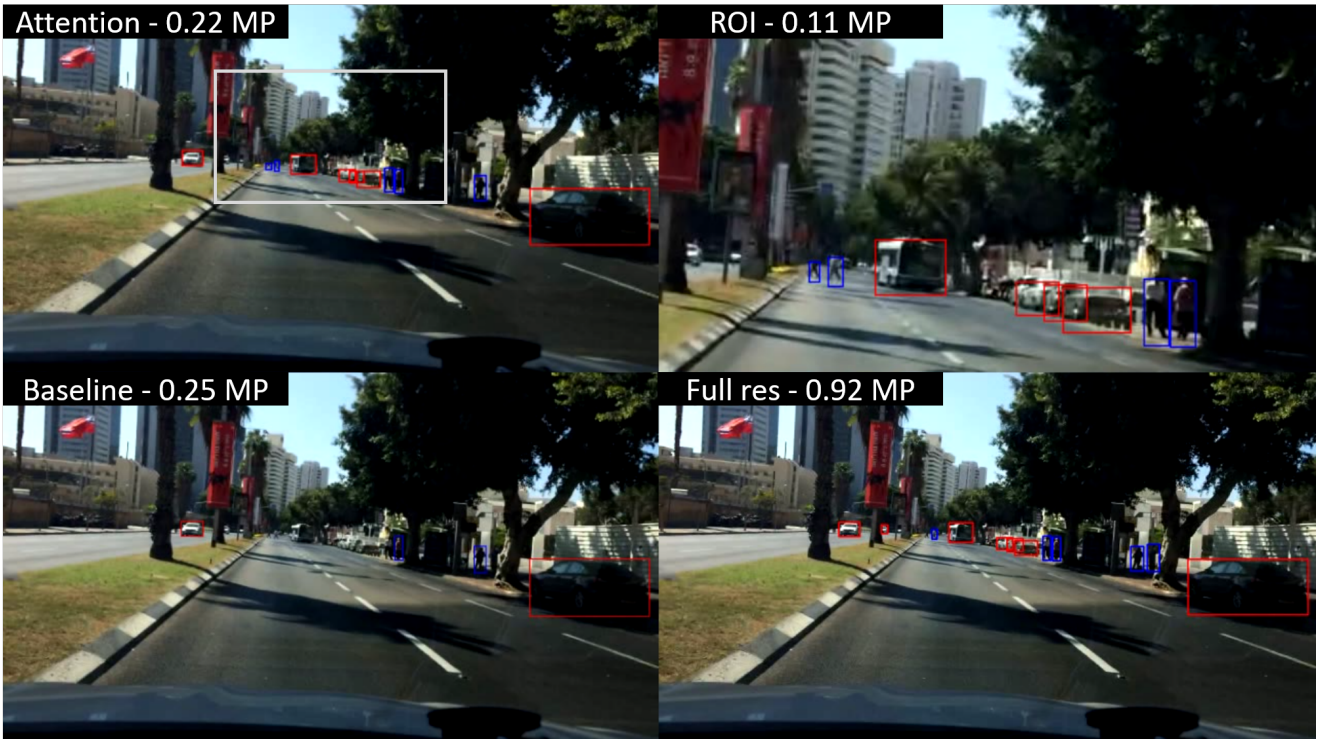

Figure 12. Illustration of evaluation results using BDD data. Notice the two distant pedestrians crossing the street, as well as the bus and parked cars along the side of the street, that were detected in the attention window, but not in the baseline view.

the vehicle class is the dominant class in the BDD dataset, the weaker performance seen in the other three classes is the reason for the less considerable performance boost seen overall in Figure 6.

The attention mechanism provided improved performance for vehicles because the DQN agent learned to place the attention window in locations that were likely to contain small, distant vehicles (i.e., near the center of the image along the horizon line). While traffic lights often appear in this region of the image, they also frequently appear above the horizon line, as well as along the left/right side of the frame when approaching intersections. Therefore, a significant performance boost is not seen for traffic lights. Similarly, bicycles and pedestrians do not often appear nearby small and distant vehicles but rather are often found on the left/right side of the frame on sidewalks. Therefore, the attention method produces diminished results for these classes. However, due to the dominance of vehicles in the BDD dataset, the overall performance (wMAP) of the attention method still outperforms the baseline despite using fewer pixels overall.

As mentioned previously in section 2.2.2, the proposed methodology is highly flexible and can be modified to account for different objectives. While the DQN trained for this study simply attempted to maximize true positive detections, the reward function can be crafted to give more significance to other classes. Furthermore, other sources of information, such as map and path planning data, can be incorporated to allow for dynamic objectives. For example, if the AV's on-board map indicates that the vehicle is approaching a traffic light, then it can employ a DQN agent trained specifically to detect traffic lights. However, if the map indicates that the vehicle is approaching a crosswalk, then the AV can employ a different DQN agent trained specifically to detect pedestrians in the AV's intended path. Incorporating dynamic objectives such as these allows the most appropriate region of interest to be identified according to the most relevant task for the AV at the moment.

As shown in Figures 6-10, the DQN-based attention scheme (dark blue) outperformed the other attention schemes. Centered attention (green) also performed favorably, considering that small, distant vehicles are often found near the center of the frame. However, it still did not perform as well as the DQN-based attention scheme, therefore, indicating that the DQN agent is, in fact, able to intelligently seek out additional objects. Random attention (light blue), on the other hand, did not perform favorably. It only slightly outperformed the low-resolution detections because, while it does provide additional resolution, there is no intelligence behind the placement of the attention window, and so the

Figure 13. Illustration of the tradeoff between higher detection accuracy within attention window and lower detection accuracy outside of attention window.

additional resolution often goes to waste.

4.3. Illustrative Example

An illustration of the detections received for the various methods is shown in Figure 12. In this figure, the HD 0.92 MP frame is shown in the bottom right quadrant, along with its detections (red for vehicles, blue for pedestrians). The baseline SD 0.25 MP frame is shown in the bottom left quadrant, along with its detections. The upper right quadrant shows the 0.11 MP region of interest (ROI) and its detections. Finally, the upper left quadrant shows the location of the attention window (white rectangle), along with the fused attention detections, which come from the 0.11 MP ROI window (top right) and the 0.11 MP low-resolution full-frame (not shown). This snapshot demonstrates how the attention mechanism provides many more small, distant detections compared to the baseline.

While the attention window gives higher detection accuracy within the region of interest, detection performance outside of the region of interest is diminished. This is expected as the proposed attention-based sensing paradigm mimics human foveal vision by providing high-resolution in one specific area, while the peripheral region is processed at a lower resolution. This tradeoff is illustrated in 13, as we see numerous vehicle and traffic light detections in the attention window, but also some missed detections outside of the region of interest along the left side of the frame. However, due to the intelligent nature in which the DQN agent places the attention window, and its ability to fixate on the most relevant region of the scene, the positive effects of this tradeoff far outweigh the negative effects.

5. Conclusion

This paper presents a novel attention-based sensing technique that leverages an end-to-end DQN-based agent to intelligently select a region of an image to place greater attention to achieve better perception performance. Due to the flexibility of the proposed method, the DQN agent can be crafted to maximize performance for various objectives. This technique presents several advantages for AD/ADAS applications, such as requiring fewer sensors, lower cost, less processing time and power consumption, improved perception capability, and enhanced HD sensors utilization. Overall, a substantial improvement in perception capability at a small cost in terms of processing time was demonstrated using the Berkeley Deep Drive dataset.

References

[1] Ana Filipa Almeida, Rui Figueiredo, Alexandre Bernardino, and José Santos-Victor. Deep networks for human visual

attention: A hybrid model using foveal vision. In *Iberian Robotics conference*, pages 117–128. Springer, 2017. 1

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014. 2

[4] Alexander V Bernstein, EV Burnaev, and ON Kachan. Reinforcement learning for computer vision and robot navigation. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 258–272. Springer, 2018. 3

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 3

[6] Sean Campbell, Niall O'Mahony, Lenka Krpalcova, Daniel Riordan, Joseph Walsh, Aidan Murphy, and Conor Ryan. Sensor technology in autonomous vehicles: A review. In *2018 29th Irish Signals and Systems Conference (ISSC)*, pages 1–4. IEEE, 2018. 1

[7] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015. 1

[8] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017. 1

[9] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 2

[10] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002. 1

[11] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. 2007. 2

[12] John M Henderson and Andrew Hollingworth. High-level scene perception. *Annual review of psychology*, 50(1):243–271, 1999. 2

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3203–3212, 2017. 2

[14] Qibin Hou, Jiang-Jiang Liu, Ming-Ming Cheng, Ali Borji, and Philip HS Torr. Three birds one stone: A general architecture for salient object segmentation, edge detection and skeleton extraction. *arXiv preprint arXiv:1803.09860*, 2018. 2

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015. 2

[16] Junsung Kim, Ragunathan Rajkumar, and Markus Jochim. Towards dependable autonomous driving vehicles: a system-level approach. *ACM SIGBED Review*, 10(1):29–32, 2013. 1

[17] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017. 2

[18] Engin Mendi and Mariofanna Milanova. Contour-based image segmentation using selective visual attention. *Journal of Software Engineering and Applications*, 3(8):796, 2010. 2

[19] Mariofanna Milanova and Engin Mendi. Attention in image sequences: Biology, computational models, and applications. In *Advances in Reasoning-Based Image Processing Intelligent Systems*, pages 147–170. Springer, 2012. 1

[20] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *arXiv preprint arXiv:1406.6247*, 2014. 2

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[22] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 3

[23] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[24] Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. Attention on attention: Architectures for visual question answering (vqa). *arXiv preprint arXiv:1803.07724*, 2018. 1

[25] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 1

[26] Ultralytics. Yolov5. https://github.com/ultralytics/yolov5, 2021. 3

[27] Michael Wunder, Michael L Littman, and Monica Babes. Classes of multiagent q-learning dynamics with epsilon-greedy exploration. In *ICML*, 2010. 3

[28] Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, Teresa Canas-Bajo, and David Whitney. Periphery-fovea multi-resolution driving model guided by human attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1767–1775, 2020. 2

[29] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1

[30] Alfred L Yarbus. *Eye movements and vision*. Springer, 2013. 1

[31] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 3