Occupancy Grid Mapping with Cognitive Plausibility for Autonomous Driving Applications

Alice Plebe
University of Trento
alice.plebe@unitn.it

Julian F. P. Kooij
Delft University of Technology
J.F.P.Kooij@tudelft.nl

Gastone Pietro Rosati Papini
University of Trento
gastone.rosatipapini@unitn.it

Mauro Da Lio
University of Trento
mauro.dalio@unitn.it

Abstract

This work investigates the validity of an occupancy grid mapping inspired by human cognition and the way humans visually perceive the environment. This query is motivated by the fact that, to date, no autonomous driving system reaches the performance of an ordinary human driver. The mechanisms behind human perception could provide cues on how to improve common techniques employed in autonomous navigation—specifically the use of occupancy grids to represent the environment. We experiment with a neural network that maps an image of the scene onto an occupancy grid representation, and we show how the model benefits from two key (and yet simple) changes: 1) a different format of occupancy grid that resembles the way the brain projects the environment into a warped representation in the cortical visual area; 2) a mechanism similar to human visual attention that filters out non-relevant information from the scene. These effective expedients can potentially be applied to any autonomous driving task requiring an abstract representation of the scenario like the occupancy grids.

1. Introduction

One of the classical cornerstones of artificial intelligence is to draw inspiration from human cognition to design similar intelligent behaviors in artificial systems [2, 15]. This tenet finds fertile ground in the field of intelligent vehicles because humans still outperform any state-of-the-art autonomous driving system. While an average human driver is expected to cause a minor accident every 10 million miles [1], the best autonomous vehicle manufacturers report an average of one disengagement every 30,000 miles [27]. Moreover, the major human causes of traffic accidents are cognitive impairments such as the influence of alcohol or drugs, tiredness, distraction, and recklessness [37]. An expert driver in normal conditions is rarely the cause of an accident. The human ability of driving is especially remarkable, given that vehicles are technological artifacts controlled by interfaces that are extraneous to the natural human motion control. Nonetheless, humans learn to drive quickly and robustly.

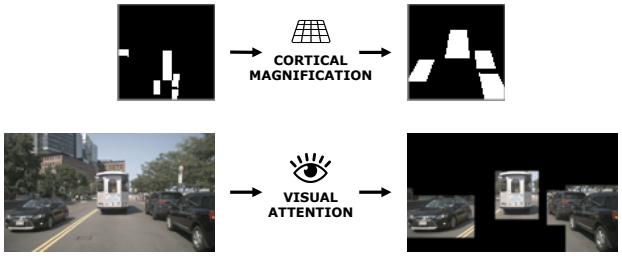


Figure 1. We propose two cognitively plausible changes to improve occupancy grid mapping: a transformation function mapping the vehicles from the visual scene into an occupancy grid that is warped in a way resembling the visual cortex (top); a technique simulating cognitive visual attention to select important information and mask non-relevant components from a cluttered scene (bottom).

We argue that the superior human driving capability suggests that the solution to achieving self-driving cars lies in the human brain itself. The cognitive mechanisms underlying the driving skill could reveal precious insights on how to design better driving agents and improve the existing techniques commonly employed in autonomous navigation. It is important to note, however, that vehicles are not biological bodies, and the hardware is not the brain. This is why most approaches to autonomous driving bear little resemblance to the cognitive processes involved during driving. It is undeniable that some of these engineering practices—like

modular decomposition—can provide highly desirable features, like reducing the complexity of single sub-modules or decoupling possible sources of failure [17]. Similarly, there are algorithms far different from brain computations that are very effective on silicon processors. Hence, it is crucial to find a compromise between well-consolidated technologies and inspiration from human neurocognition.

This work investigates the integration of cognitive plausibility in the task of occupancy grid mapping, i.e., generating from a visual scene a representation of the vehicle's surrounding environment. We apply cues derived from known brain mechanisms to improve the format of occupancy representation and to simplify the visual information coming from the environment. Our contributions are the following:

1. We modify the occupancy mapping function such that the occupancy grid is not uniformly spaced anymore, but warped in a way that mimics human *cortical magnification* [18], which is the space warping applied by the primary cortical visual area to enlarge the central part of the scene with respect to the peripheral parts;

2. Instead of considering the entire visual scene, we suppress the parts of the image that do not contain crucial elements (vehicles); in this way, we simulate the human *visual attention* mechanisms [10, 20, 26, 35], which deal with cluttered visual scenes by selecting important information and by filtering out non-relevant information.

Note that the cognitive mechanism of visual attention considered here is very different from the popular notion of attention in deep learning—we will clarify this aspect in Section 2.2.

This work provides an example of how key neurocognitive principles can be translated into simple expedients to be integrated into existing autonomous navigation approaches, obtaining valid improvements without much computational effort. Here we show the case of occupancy grid mapping, but the idea of cognitive plausibility has been successfully investigated also in other autonomous driving tasks [6, 30–32].

The rest of the paper is organized as follows. Section 2 summarizes the existing methods of occupancy grid mapping, and it illustrates the different accounts of attention mechanism in the literature. Section 3 describes our two contributions: a formulation of occupancy grids that conciliates between mathematical efficiency and cognitive plausibility; a mechanism that emulates the beneficial effects of visual attention in perception of driving scenarios. The section also illustrates the model implementation. Section 4 presents the dataset, the evaluation metrics, and the results on different combinations of input and output formats along with a comparison with the related works. Lastly is Section 5 which presents the conclusions.

2. Related Works

2.1. Occupancy Grid Mapping

One of the most established formats of abstract representation of driving environments is the *occupancy grids*, first proposed in the field of robotic perception and navigation [12]. An occupancy grid G is a uniformly spaced 2D lattice of binary elements $g_{i,j}$ each representing the presence of an obstacle at the corresponding patch $A_{i,j}$ of the continuous world space:

$$g_{i,j} = \begin{cases} 1 & \text{if } \exists\, o_k \mid \left(X^{(k)}, Z^{(k)}\right) \in A_{i,j} \\ 0 & \text{otherwise} \end{cases}, \qquad (1)$$

where o_k is an object with center $\left(X^{(k)}, Y^{(k)}, Z^{(k)}\right)$ located in the environment (considering the Z-axis pointing to the travel direction of the ego car, the X-axis pointing to the right of the ego vehicle, and the Y-axis perpendicular to the ground plane pointing upwards). The patch $A_{i,j}$ is the area on the ground plane defined by a neighborhood of the point (X_i, Z_j) of width $[\Delta X, \Delta Z]$.

An occupancy grid can be also interpreted as a binary image. Assuming the origin of the coordinate system of the image space in the top-left corner and the bottom-right corner having coordinates $(W-1, H-1)$, the transformation of a world point (X_i, Z_j) into image coordinates (i, j) is the following:

$$i = \frac{W}{2} + \frac{X_i}{\Delta X}, \qquad (2)$$

$$j = H - \frac{Z_j - \widetilde{Z}}{\Delta Z}, \qquad (3)$$

where the parameter $\widetilde{Z}$ represents the longitudinal length of the "blind zone" of the ego camera, i.e., the area immediately in front of the ego vehicle which lies outside the field of view of the camera. Since the information about the obstacles o_k is the result of measurements affected by uncertainty, the values of $g_{i,j}$ are usually probabilities rather than simple binary values.

Probabilistic occupancy grids have become popular also outside the domain of robotics. They are frequently adopted in high-level modules of autonomous driving systems—a recent overview of their applications can be found in [25]. A precious feature of occupancy grids is the predisposition to be processed by artificial neural networks. Because of their 2D matricial format, they combine effectively with convolutional neural networks. Moreover, the values in a grid cell can span multiple channels and include additional information, such as the semantic of the obstacle or its velocity.

Most of the works on occupancy grid mapping leverage range sensors like LIDARs and radars, which provide the depth information to easily generate a *bird-eye's view*

(BEV) of the scene. Some works adopt other visual sensors that include depth information, such as RGB-D cameras [16] and stereo cameras [22], while other works fuse multiple sensors, e.g. cameras and LIDARs [13, 28].

There are a few works that map occupancy grids directly from a monocular camera, as in our approach. Lu et al. [23] use a variational encoder-decoder to produce a top-down semantic representation of the scene. Their work, however, focuses on detecting the drivable areas rather than the vehicles, and it puts special emphasis on training the model with weak ground truth. Mani et al. [24] predict both the BEV layout of the scene and the vehicles location. They propose to "hallucinate" plausible completions for occluded parts by leveraging adversarial feature learning. Roddick & Cipolla [36] too generate top-down maps capturing both the road layout and the traffic participants. They learn an implicit mapping from the image plane to the BEV plane using the information on the camera geometry. Philion & Fidler [29] produce BEVs that include vehicles, drivable areas, and lane boundaries, leveraging multi-view camera data coming from a full camera rig covering 6 points of view. Can et al. [5] predict semantic top-down maps with several dynamic and static classes. They exploit additional information coming from temporal aggregation and feed the model sequences of frames as input.

In contrast with these works, our approach leverage exclusively single images of the monocular camera—without the need of camera geometry information, temporal data, or multiple cameras—and focuses on mapping the occupancy of the vehicles, with emphasis on the accuracy of vehicles that are close to the ego car.

2.2. Attention Mechanisms

The term "attention" has recently become very popular in the computer vision community, with the rise of the so-called *self-attention* vision models. It is important to highlight that these models have little to do with human visual attention. As a consequence, our proposed mechanism imitating natural attention is profoundly different from the common account of self-attention mechanism in deep learning.

The deep learning adaptation of "attention" started within the field of Natural Language Processing [38], with the aim to capture long-range dependencies. Self-attention makes it possible to model long-distance interactions in neural sequence transduction models, even without the need of recurrent layers [38]. By applying a similar approach to tasks like image recognition and video classification [39], self-attention allows vision models to augment—or entirely replace [33]—the convolution operations and capture long-range dependencies in visual data.

Natural visual attention is not about spatial or temporal long-distance interactions. Although the term encompasses a set of cognitive mechanisms, it mainly refers to the close relation between saccadic eye movements and covert orienting of visual spatial attention [10, 20, 26, 35]. Saccades are rapid movements of the eyes that change the point of foveal fixation. The saccade location proves to be where the attention of a person is mostly directed; in fact, one cannot move the eyes to one location and attend to a different one.

The saccadic system has inspired similar mechanisms in computer vision. For example, De Souza et al. [7] implement a variant of the retinal log-polar transform for a neural network to detect traffic signs using a small dataset. Note that there are also computational models that implement aspects of natural visual attention accurately [8, 9, 19, 21]. However, the purpose of these models is to help investigate the neurocomputational basis of visual attention, and they are not aimed at engineering applications requiring efficiency and high performance.

We propose an attention mechanism that follows the cognitive account of attention. Given that the task at hand is mapping the location of vehicles in the occupancy grids, the model needs to direct its "attention" towards the spatial locations of the vehicles in the input image. We achieve this by suppressing the visual information not concerning vehicles.

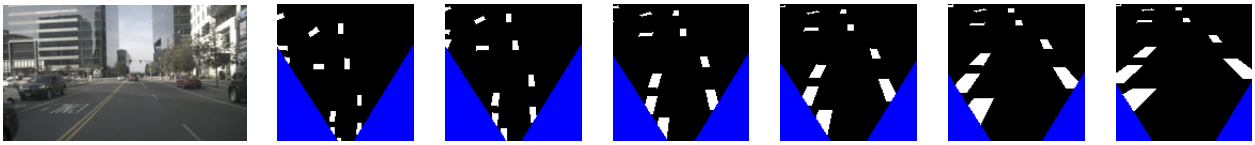
3. Methodology

This work deals with the task of occupancy grid mapping: given an image of a traffic scene, predict an occupancy grid of the surrounding vehicles. To realize this task, we have implemented a deep convolutional encoder-decoder. The objective of this work is to show that cognitive principles can help design mechanisms to improve prediction of occupancy grids. We propose two cognitively plausible mechanisms—the first revising the format of the output, the second pre-processing the input.

3.1. Cognitively Plausible Occupancy Grids

Two core aspects of occupancy grids go strongly against cognitive plausibility: the point of view of the image and the uniform tessellation of the world space. Firstly, an occupancy grid—as conventionally used in the context of navigation—corresponds to a BEV or a top-down view of the world space. This kind of orthographic view is clearly impossible for a human driver. Secondly, in the human visual system, the retinal space is never represented uniformly. In the primary cortical visual area, the space warping with respect to the eye view is known as *cortical magnification* [18]. This warping enlarges the central space of the scene at the expense of the peripheral areas, and it is typically described with the polar-log transformation [3, 11].

We propose an occupancy grid representation that better resembles the human perception system while preserving



original $\omega = 50$ $\omega = 10$ $\omega = 5$ $\omega = 2$ $\omega = 1$

Figure 2. Warped occupancy grids representing a visual scene (leftmost image) using different values of warping.

the undeniable computational advantages of the simple matricial structure of the grids. By applying a similar warping to the occupancy grid, we magnify close objects and reduce the size of distant objects. Just like with cortical magnification, in the "warped occupancy grid", the more relevant an object is the more it is represented in detail. In the case of driving, the relevance of an object depends mainly on the time required by the ego car to reach it: closer objects are more crucial, and they require more precision. For this reason, we exploit the idea of cortical magnification in the longitudinal distance only, rather than considering a full polar-log transformation in all dimensions.

We define the warping transformation as a logarithmic transformation in the longitudinal dimension (Z-axis) and as a linear transformation in the lateral dimension (X-axis). In this way, every element of the warped occupancy grid corresponds to a square patch of the world space. The longitudinal warping transformation is the following:

$$w(Z) = \log(Z + \omega) - \log(\omega), \tag{4}$$

where ω is the constant defining the amount of warping, with the maximum deformation at $\omega = 1.0$ and no deformation for $\omega \to +\infty$. Fig. 2 gives an example of how ω influences the appearance of a warped occupancy grid. The transformation of a world point into image coordinates in the warped occupancy grid is the following:

$$i = \frac{W}{2} + \frac{H\Delta Z}{w(H\Delta Z)} \frac{w(Z_j)}{Z_j} \frac{X_i}{\Delta X}, \tag{5}$$

$$j = H - H\frac{w(Z_j) - w(\widetilde{Z})}{w(H\Delta Z)}. \tag{6}$$

This transformation magnifies the objects that are closer to the ego camera, in contrast with the linear occupancy grid transformation—eqs. (2) and (3)—in which objects have constant size in every point of the image space. The warped occupancy grid appears more similar to what a human perceives while driving, having a comparable perspective and point of view of the scene, and giving more importance to the objects in the foreground. At the same time, the warped occupancy grid still corresponds to conventional coordinates in the real world, and it can be potentially used by control systems for navigation without any additional step.

Figure 3. Examples of the attention mechanism applied in three scenarios: on the first row the original images, on the second row the results of "focusing the attention" on the vehicles.

3.2. Cognitively Plausible Visual Attention

There are many discrepancies between human perception and artificial perception, especially in the context of driving. One above all concerns dealing with cluttered visual scenes. Normally in artificial perception, the system acquires images covering the whole scenario populated by numerous objects, which may be relevant or negligible. Then, elaborate processes analyze the entire images to attempt to recognize the relevant objects and locate them in the world geometry.

In human perception, a set of cognitive operations deals with complex visual scenes by selecting important information and by filtering out irrelevant information. For example, a person driving through the countryside does not need to first classify the kind of fruits hanging from the trees by the roadside to understand they are not relevant information—the driver instinctively directs the attention to the road and the cars ahead. This cognitive process takes the name of *visual attention*, and it is one of the most studied topics in visual science [10, 20, 26, 35].

We propose a mechanism imitating the role of visual attention to reduce the computational complexity of the neural network. When dealing with traffic scenes, the salient parts of the scenes are mainly the vehicles. The attention should be directed to them, rather than to the surrounding buildings, foliage, or sky—especially in a task like occupancy grid mapping. Hence, we simulate the effect of attention simply by suppressing the areas of the image where there are no vehicles, before feeding the image to the model. Fig. 3 shows the result of this preprocessing. We use YOLO-v3 [34] to detect in the image the 2D bounding boxes containing vehicles and suppress the pixels outside

these regions.

3.3. Model Architecture

The proposed model is a deep convolutional neural network with a encoder-decoder architecture. The input of the model is a single image of 800×450 pixels representing the traffic scene. The output is a graylevel image of 128×128 pixels that represents the occupancy grid describing the vehicles location in the scene. The encoder is composed of a stack of seven convolutions followed by two fully-connected layers, ending into a latent space of 256 neurons. The decoder consists of a single fully-connected layer followed by a stack of five deconvolutions. The loss function used to train the model is the binary cross-entropy.

We have implemented different versions of the model varying the formats of input and output. The two formats of input are the original frame coming from the video stream of the ego camera (FRM), and the frame processed with the attention mechanism (ATT). The output formats are the standard uniformly spaced occupancy grid (OCC), and the warped occupancy grid (WRP). Moreover, we have tested two versions of WRP with different coefficients of warping: $\texttt{WRP}_1$ uses $\omega = 1$, and $\texttt{WRP}_2$ uses $\omega = 2$. Ultimately, there are six variations of the model corresponding to the possible combinations of input and output.

4. Results

4.1. Dataset

The dataset adopted in this work is *nuScenes* [4], developed by the company Motional (formerly known as nuTonomy). The dataset is organized into 1000 video sequences (of which 850 sequences include LIDAR data and annotations) featuring a considerable variety of environments, illuminations, and weather conditions. For a swifter training and testing process, we reduce the dataset to 200 video sequences (∼40,000 frames). The reduced dataset still preserves a good variety of driving scenarios, but it excludes the sequences in which there are no vehicles for most of the time. We randomly allocate 140 of the selected sequences to the training set, 30 to the validation, and 30 to the test set.

To generate the ground truth data for the occupancy grids, we map the 3D bounding box annotations onto binary images of 128×128 pixels using the two transformation functions defined in equations (2), (3) for the standard occupancy grid (OCC) and (5), (6) for the warped occupancy grid (WRP). We set $\Delta X = \Delta Z = 0.5$ m so that every pixel of the occupancy grid corresponds to a square of 0.25 m^2 of the world space, and $\widetilde{Z} = 3.5$ m. Hence, the maximum distance represented in the occupancy grids (both uniform and warped) is 67.5 m.

To implement the attention mechanism, we use YOLO as "out-of-the-box" object detector. We find YOLO does not need fine-tuning on nuScenes because the predicted 2D bounding boxes appear consistent with the ground truth of the dataset.

4.2. Evaluation Metrics

We consider three metrics to evaluate how the cognitive variations improve the performance of the model. The metrics are the intersection over union (IoU), the average precision (AP), and the distance between centroids in the world space. The IoU is computed between the entire target and predicted occupancy grid images. Instead, the AP and the centroids are computed with respect to the connected regions extracted from the occupancy grids. The matches between connected regions of target and prediction are determined using the greedy algorithm adopted from KITTI [14]. We set a binarization threshold of $\theta = 0.4$ and $N_\Theta = 40$ recall levels for the AP computation.

We are aware that IoU and AP might favor WRP with respect to OCC, because of the different spaces in which the metrics are computed. That is why we have included the metric measuring the distance between centroids. This metric is not usually adopted in works on occupancy grid mapping, but it is computed in terms of the original coordinates of the world space, so it provides a fair evaluation in this context. In fact, the results in Section 4.3 will show similar trends in all three metrics.

To better assess how the warped occupancy grid improves the prediction of closer vehicles, we also compute the metrics separately in three different classes of depth in the world space. The occupancy grids and the related connected regions are partitioned into a close range (CLS) for depth < 15 m, a far range (FAR) for depth > 30 m, and a middle range (MID) in between.

4.3. Results

Table 1 shows the results of the proposed model comparing the six combinations of input and output formats described in Section 3.3. All model variations share the same architecture and hyperparameters, and they have been trained for 200 epochs. The results are grouped by evaluation metric, and each group is in turn divided into the four classes of depth range described in Section 4.2 (ALL refers to the entire range).

The baseline FRM-OCC maps the image frame into a standard uniformly spaced occupancy grid. The models FRM-WRP change the output format as they map the frame into the warped occupancy grids emulating cortical magnification. The model ATT-OCC employs the cognitive mechanism on the input instead, masking the areas of the image without vehicles. The models ATT-WRP combine both methods. For the warped occupancy grid, the table reports separate results for the two warping factors, namely $\texttt{WRP}_1$ and $\texttt{WRP}_2$.

It is possible to identify some patterns that seem to occur

	IoU ↑				Average Precision ↑				Distance Centroids ↓			
	ALL	CLS	MID	FAR	ALL	CLS	MID	FAR	ALL	CLS	MID	FAR
FRM – OCC	16.4	42.0	26.0	12.2	0.077	0.195	0.103	0.022	1.74	1.41	1.80	2.11
FRM – WRP_1	30.3	47.0	30.9	8.7	0.096	0.254	0.085	0.005	1.51	1.24	1.62	5.18
FRM – WRP_2	31.3	46.6	33.4	9.1	0.089	0.231	0.096	0.009	1.15	1.04	1.36	4.25
ATT – OCC	20.3	48.5	30.1	14.9	0.113	0.319	0.154	**0.034**	1.58	1.39	1.61	**2.06**
ATT – WRP_1	32.9	**53.9**	34.1	13.6	0.142	**0.416**	0.148	0.021	1.15	**0.95**	1.39	4.56
ATT – WRP_2	**34.0**	53.7	**38.1**	**15.2**	**0.144**	0.414	**0.175**	0.027	**0.99**	0.98	**0.97**	3.96

Table 1. Comparison of the formats of input and output over the same architecture. The scores are computed on the entire output and on partitions of the output space based on the distance from the ego car.

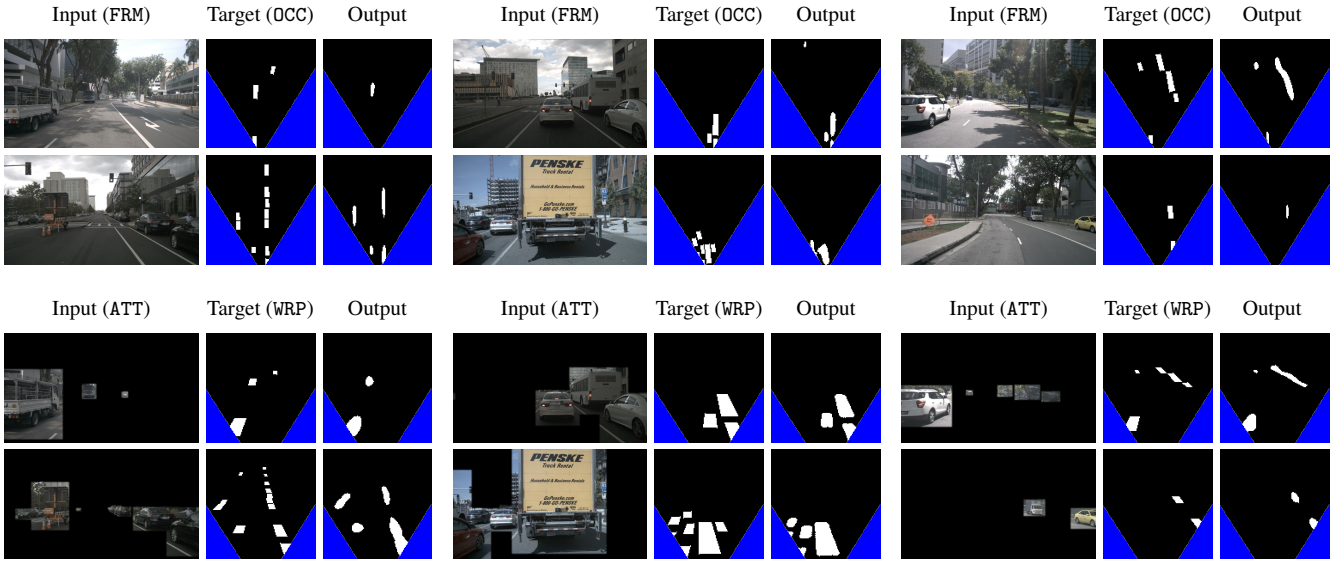


Figure 4. Visual results on six test samples using the baseline (FRM-OCC) and the model with cognitive plausibility (ATT-WRP_2). The predicted output is displayed with a binarization threshold of 0.4.

consistently. Firstly, the models with input ATT outperform the models with the corresponding output and input FRM. This happens consistently in all the metrics. It can be deduced that the proposed attention mechanism benefits the training of the neural network. Secondly, the models that perform better in the close and middle ranges (CLS and MID) are the models with ouput WRP. This happens, again, in all three metrics. The superiority of WRP in the middle/close range is coherent with the idea that the warping mimics the effect of cortical magnification and improves the accuracy of the vehicles in the proximity of the ego car. Moreover, while the strongest warping factor WRP_1 achieves the best CLS scores, the model with WRP_2 performs better in general, which is indicated by the higher ALL scores in all metrics. Lastly, the models with WRP manifest a larger gap between the scores in CLS and FAR, with respect to the model using OCC as output. This is, once again, consistent with the fact that the standard occupancy grid is uniformly spaced. Hence, a car close to the camera (which occupies a large portion of the image frame) and a very distant car (displayed in few pixels of the frame) occupy a similar number of grid elements and are predicted with similar accuracy. Fig. 4 shows visual results on six test samples using the baseline FRM-OCC (top of figure) and the best model with cognitive plausibility ATT-WRP_2 (bottom).

Table 2 presents a comparison with the other relevant approaches mapping visual input into occupancy grids reviewed in Section 2.1. We have selected the work that have trained their models on nuScenes. Table 2 reports the IoU scores provided in the original papers. Although all methods employ nuScenes, the video sequences in the training and test sets are not the same in all cases. Moreover, it should be recalled that we train our models on a reduced number of nuScenes sequences, as described in Section 4.1. Hence, this comparison is to be considered in broad terms.

The combination of the two proposed mechanisms (ATT-WRP_2) drastically improve the performance of the baseline (FRM-OCC). These cognitive-inspired mechanisms are easy

	IoU
Lu et al. [23]	8.8
Roddick & Cipolla [36]	24.7
Philion & Fidler [29]	32.1
Can et al. [5]	36.0
Ours (FRM-OCC)	16.4
Ours (ATT-WRP$_2$)	34.0

Table 2. IoU scores (%) on the *nuScenes* dataset. We report the scores from the original papers.

to implement and employ a light architecture easy to inspect. The related approaches adopt heavier neural networks which exploit additional information, like temporal data [5], multiple cameras [29], and camera geometry [36]. On the other hand, these works do not focus specifically on the vehicles: they also predict semantic maps of the road and the drivable areas. Therefore, the results in Table 2 are, once again, an approximate comparison. Nonetheless, while the baseline (FRM-OCC) obtains limited performance, the model with cognitive plausibility (ATT-WRP$_2$) is competitive compared to most of the state-of-the-art methods.

5. Conclusions

We have investigated the benefit of integrating cognitive plausibility into autonomous driving tasks, specifically in the case of occupancy grid mapping. Our method improves the baseline approach to occupancy grid mapping by introducing two cognitively plausible changes. First, we have simulated the cognitive mechanism of visual attention by suppressing in the visual input the areas of the image that do not contain vehicles. Second, we have mimicked human cortical magnification by warping the occupancy grid in order to magnify close vehicles and improve their accuracy. The two improvements require minimum implementation effort compared to other related state-ot-the-art approaches, obtaining competitive results on the *nuScenes* dataset. Our results indicate the value of taking human cognition as a source of inspiration—an idea that could be explored in different tasks for intelligent vehicles.

There is a fundamental discrepancy between our approach and the cognitive mechanisms involved during driving: the lack of dynamic information. Natural vision is inherently dynamic. Humans do not perceive static independent "frames"; on the contrary, they rely heavily on temporal dynamics of the environment. The current work does not leverage information about the movements of the objects. Hence, a promising future development could be a way to integrate dynamic information in the model to better support the cognitive inspiration.

References

[1] U.S. National Highway Traffic Safety Administration. Traffic safety facts annual report tables, 2018. 1

[2] Margaret Boden. *AI: Its nature and future*. Oxford University Press, Oxford (UK), 2016. 1

[3] Richard Born, Alexander R. Trott, and Till S. Hartmann. Cortical magnification plus cortical plasticity equals vision? *Vision Research*, 111:161–169, 2015. 3

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 5

[5] Yigit Baran Can, Alexander Liniger, Ozan Unal, Danda Paudel, and Luc Van Gool. Understanding bird's-eye view semantic hd-maps using an onboard monocular camera. *arXiv preprint arXiv:2012.03040*, 2020. 3, 7

[6] Mauro Da Lio, Riccardo Donà, Gastone Pietro Rosati Papini, and Kevin Gurney. Agent architecture for adaptive behaviors in autonomous driving. *IEEE Access*, 8:154906–154923, 2020. 2

[7] Alberto F De Souza, Cayo Fontana, Filipe Mutz, Tiago Alves de Oliveira, Mariella Berger, Avelino Forechi, Jorcy de Oliveira Neto, Edilson de Aguiar, and Claudine Badue. Traffic sign detection with vg-ram weightless neural networks. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2013. 3

[8] Gustavo Deco. Biased competition mechanisms for visual attention in a multimodular neurodynamical system. In Stefan Wermter, Jim Austin, and David Willshaw, editors, *Emergent neural computational architectures based on neuroscience: towards neuroscience-inspired computing*, pages 114–126. Springer-Verlag, Berlin, 2001. 3

[9] Gustavo Deco and Edmund Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44:621–642, 2004. 3

[10] R. Desimone. Neural circuits for visual attention in the primate brain. In G. A. Carpenter and S. Grossberg, editors, *Neural Networks for Vision and Image Processing*. MIT Press, Cambridge (MA), 1992. 2, 3, 4

[11] Robert O. Duncan and Geoffrey M. Boynton. Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron*, 38:659–671, 2003. 3

[12] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22:46–57, 1989. 2

[13] Özgür Erkent, Christian Wolf, Christian Laugier, David Sierra Gonzalez, and Victor Romero Cano. Semantic grid estimation with a hybrid Bayesian and deep neural network approach. In *International Conference on Intelligent Robots and Systems*, pages 1–8, 2018. 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 5

[15] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95:245–258, 2017. 1

[16] Marian Himstedt and Erik Maehle. Online semantic mapping of logistic environments using rgb-d cameras. *International Journal of Advanced Robotic Systems*, 14(4):1729881417720781, 2017. 3

[17] Yu Huang and Yue Chen. Autonomous driving with deep learning: A survey of state-of-art technologies. *arXiv*, abs/2006.06091, 2020. 2

[18] David Hubel and Torsten Wiesel. Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *Journal of Comparative Neurology*, 158:295–305, 1974. 2, 3

[19] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001. 3

[20] Nancy Kanwisher and Ewa Wojciulik. Visual attention: insights from brain imaging. *Nature Reviews Neuroscience*, 1:3310–3318, 2000. 2, 3, 4

[21] Sofia Krasovskaya and W. Joseph MacInnes. Salience models: A computational cognitive neuroscience review. *Vision*, 3:vision3040056, 2019. 3

[22] You Li and Yassine Ruichek. Occupancy grid mapping in urban environments from a moving on-board stereo-vision system. *Sensors*, 14(6):10454–10478, 2014. 3

[23] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. *IEEE Robotics and Automation Letters*, 4:445–452, 2019. 3, 7

[24] Kaustubh Mani, Swapnil Daga, Shubhika Garg, Sai Shankar Narasimhan, Madhava Krishna, and Krishna Murthy Jatavallabhula. Monolayout: Amodal scene layout from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1689–1697, 2020. 3

[25] Sajjad Mozaffari, Omar Y. Al-Jarrah, Mehrdad Dianati, Paul Jennings, and Alexandros Mouzakitis. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, early access:1–15, 2020. 2

[26] Scott O. Murray and Sheng He. Contrast invariance in the human lateral occipital complex depends on attention. *Cerebral Cortex*, 16:606–611, 2001. 2, 3, 4

[27] California's Department of Motor Vehicles. Disengagement report, 2021. 1

[28] Sang-Il Oh and Hang-Bong Kang. Fast occupancy grid filtering using grid cell clusters from lidar and stereo vision sensor data. *IEEE Sensors Journal*, 16(19):7258–7266, 2016. 3

[29] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 3, 7

[30] Alice Plebe and Mauro Da Lio. On the road with 16 neurons: Towards interpretable and manipulable latent representations for visual predictions in driving scenarios. *IEEE Access*, 8:179716–179734, 2020. 2

[31] Alice Plebe and Mauro Da Lio. Neurocognitive-inspired approach for visual perception in autonomous driving. In *Communications in Computer and Information Science*, volume 1217, pages 113–134. Springer, Cham, 2021. 2

[32] Alice Plebe, Riccardo Donà, Gastone Pietro Rosati Papini, and Mauro Da Lio. Mental imagery for intelligent vehicles. In *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS)*, pages 43–51. Science and Technology Publications, 2019. 2

[33] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *33rd Conference on Neural Information Processing Systems*, 2019. 3

[34] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv*, abs/1804.02767, 2018. 4

[35] John H. Reynolds, Leonardo Chelazzi, and Robert Desimone. Competitive mechanisms subserve attention in macaque areas V2 and V4. *Nature Neuroscience*, 2:1019–1025, 1999. 2, 3, 4

[36] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11138–11147, 2020. 3, 7

[37] Santokh Singh. Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. Technical Report DOT HS 812 115, National Highway Traffic Safety Administration, Washington (DC), 2015. 1

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017. 3

[39] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 3