Semantic-Based Active Perception for Humanoid Visual Tasks with Foveal Sensors

João Luzio, Alexandre Bernardino, Plinio Moreno
Instituto Superior Técnico
Lisbon, Portugal
joaoluzio14@tecnico.ulisboa.pt, {alex, plinio}@isr.tecnico.ulisboa.pt

***Abstract*—The aim of this work is to establish how accurately a recent semantic-based foveal active perception model is able to complete visual tasks that are regularly performed by humans, namely, scene exploration and visual search. This model exploits the ability of current object detectors to localize and classify a large number of object classes and to update a semantic description of a scene across multiple fixations. It has been used previously in scene exploration tasks. In this paper, we revisit the model and extend its application to visual search tasks. To illustrate the benefits of using semantic information in scene exploration and visual search tasks, we compare its performance against traditional saliency-based models. In the task of scene exploration, the semantic-based method demonstrates superior performance compared to the traditional saliency-based model in accurately representing the semantic information present in the visual scene. In visual search experiments, searching for instances of a target class in a visual field containing multiple distractors shows superior performance compared to the saliency-driven model and a random gaze selection algorithm. Our results demonstrate that semantic information, from the top-down, influences visual exploration and search tasks significantly, suggesting a potential area of research for integrating it with traditional bottom-up cues.**

***Index Terms*—active perception, foveal vision, visual search, object detection, visual saliency**

I. Introduction

Foveal vision [1] and active perception [2] are the foundation of human visual cognition [3]. Foveal vision reduces the amount of information to process during each gaze fixation, while active perception sequentially changes the gaze direction to the most promising regions of the visual field. Foveal vision is a means of improving the efficiency of brain computations, as, at each time, the brain has to process a smaller amount of information. In addition, foveal vision reduces the amount of information at the periphery of the visual field [1] which reduces the saliency of the objects at those locations, e.g. [3] (see Fig. 1). This challenges modern computer vision methods and models [4] in terms of their ability to detect (localize and classify) such blurred objects.

Traditional cognitive models from psychology [5] state that the guiding mechanism for human visual attention depends on a range of distinct types of conspicuity features, which are combined in a saliency map that highlights regions of interest. Modern approaches consider visual models that can extract bottom-up [6], top-down [7], or both types of features [8] to actively perform typical human visual tasks.

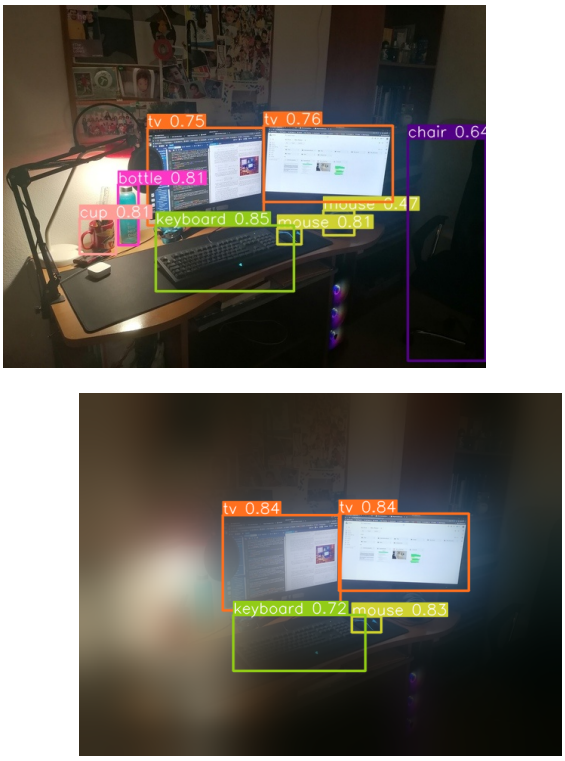

Fig. 1. Example of a regular image (top) that is artificially foveated (bottom) to emulate a visual field, together with the object predictions, outputted by an object detector, represented by their bounding-boxes and confidence scores.

Recent work, developed by Dias et al. [9], explores the possibility of combining the rich semantic information provided by modern object detectors with active perception, to perform scene exploration in foveal images. When exploring a scene, the goal is to accurately map the semantic content available in a visual field through successive premeditated gaze shifts. The main challenges associated with this humanoid task are the ability to cope with the reduced resolution of the visual information available in the peripheral region and the accumulation of semantic information in successive gaze shifts [10] to improve the agent's belief in the presence of objects from a set of known classes. Both challenges are addressed using Bayesian methods [11] and proper characterization of the uncertainty that is correlated with foveal image degradation.

In [9] the semantic data generated by the object detectors are

interpreted as Dirichlet random variables [12], which enable proper data fusion [13] and calibration [14]. Active perception methods are then formulated as determining gaze locations that minimize some measure of uncertainty. In this work, we consider applying the same methodology [9] to another visual task performed by humans, namely visual search [15], where the objective is to set the gaze directly upon a region that contains an instance of a pre-defined target class. For this reason, the relevance of gaze selection is amplified given the greedy target-oriented nature of this cognition-related task.

In summary, the objective of this work is to establish how well the novel semantic-based active perception model [9] is able to complete visual tasks that humans routinely perform. For this purpose, we compare its accuracy with a biologically inspired attention model [6], based on studies of human visual cognition [5], which attempts to mimic both the behavior and the neuronal architecture of the primate visual system. Following an appropriate experimental setup, we confirm that the semantic model is applicable in the context of visual search. More specifically, we can conclude that the semantic model significantly outperforms both a random gaze selection algorithm and the selected bottom-up saliency-based attention model [6] when considering the experimental results obtained while performing both visual search and exploration tasks.

The outline of this document is as follows. In Section II we present a compilation of key concepts that are fundamental to our work. Then, in Section III we describe the complete methodological apparatus that involves the semantic-based foveal active perception, to be applied during the experimental phase. In Section IV, we present an overview of the experimental setup along with a comparison of the most relevant results obtained during the experiments, upon completion of multiple visual search tasks. A comparison of different approaches to active perception is established. Finally, in Section V we highlight the most important contributions of this work and present the final remarks.

II. Background and Related Work

A. Humanoid Visual System

Despite being often compared to a photographic camera, the human eye's processing capacity across the visual field is not homogeneous. Several anatomical properties of the eye are, for example, correlated with the presence of gaps [1] in sensory information. Due to these anatomical properties, the processing of visual signals varies quite dramatically across the visual field. Therefore, it is important to distinguish between the center of the visual field, known as the fovea, and an outer region, known as the periphery. Foveal vision ensures maximum acuity and contrast sensitivity in a small area around the gaze position [1], while peripheral vision allows for a large field of view, although it presents lower resolution and contrast sensitivity, as well as higher object positional [3].

A classical approach to generate artificially foveated images from regular Cartesian images is to use methods based on log-polar transformations [16] to generate cortical maps. A cortical map is a topographic representation of sensory or motor information in the brain, specifically found in the cerebral cortex. A log-polar space is the most appropriate approach to generate cortical images [17], modeling with reasonable fidelity the mapping observed in the primate visual cortex. Other approaches apply Cartesian foveal geometry [19] to avoid the nonlinearities [16] associated with log-polar spaces. In our methodology, we use an artificial foveal system proposed by Almeida et al. [20] to foveate regular Cartesian images. This system operates on the basis of a combination of Gaussian and Laplacian pyramids, taking advantage of space-variant low-pass spectral filters to introduce different levels of blur on the distinct regions of the image, precisely as depicted in Fig. 2. One advantage of this foveation model is that it is fast [21], making it applicable in a real-time context.

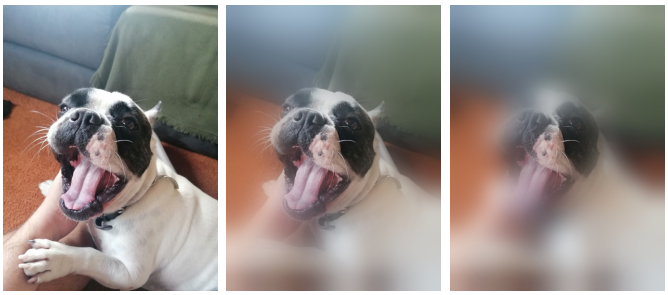

Fig. 2. Example of a Cartesian image (left) to which the artificial foveal system, proposed by Almeida et al. [20], has been applied. There are presented two foveal images, with wider (center) and narrower (right) foveal dimensions.

B. Semantic Object Detection

In general, object detection models aim at classifying existing objects in any provided image, typically identified by a rectangular box (bounding-box) and an associated vector of scores, typically normalized to be interpreted as class posterior probabilities. As described in [4], there are three main stages in traditional object detection models: selection of informative regions, extraction of features, and classification of objects. Several approaches follow this pipeline when solving object detection, each of them unique in some particular aspect. Object detection frameworks can mainly be divided into two types: The first approach (two-stage frameworks) follows the traditional pipeline of stages and consists of first generating region proposals and then classifying each proposal into the relevant object categories (e.g. R-CNN [22]). In the second approach (one-stage frameworks), object detection is viewed as a regression or classification problem, where both categories and locations are determined directly through the use of a unified framework (e.g. SSD [23], YOLO [24], [25]). These frameworks have been propelled by the overwhelming power of Convolutional Neural Networks (CNNs) given their ability to process features that are hierarchy extracted from images. Moreover, while traditional object detection methods (e.g. [22]) involve separate stages for region proposal and object classification [4], more modern transformer-based approaches (e.g. EVA [26], DETR [27]) aim to perform both tasks in a

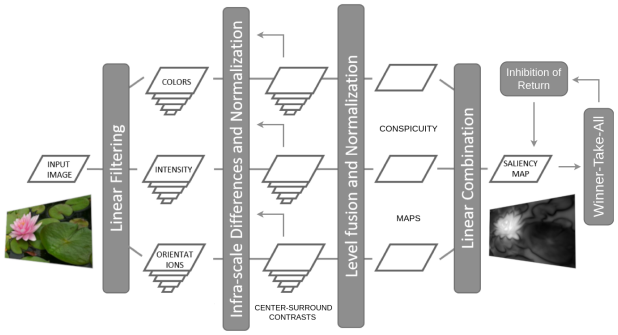

Fig. 3. Simplified version of the Itti-Koch attention system [5], inspired by both the neuronal architecture and behavior of the early primate visual system.

single step, leveraging the attention mechanism in transformers for capturing contextual information across the image.

C. Human Cognitive Mechanisms

Biological organisms use selective visual attention mechanisms to process single portions of the available visual information while disregarding the remainder. This enables these organisms to efficiently perceive and handle the limited neural resources in the brain, which are allocated to vision. Furthermore, visual attention covers all factors that influence information selection mechanisms [3], whether guided by visual stimuli, which we refer to as bottom-up features, or by task-related expectations, denominated as top-down features. The Attentional Engagement Theory (AET), also known as Similarity Theory, asserts that stimulus objects are represented at a level of perceptual description, where top-down object representations [15] compete with each other to enter visual short-term memory. These representations enhance the target-distractor dissimilarities when the competition is biased toward certain target objects, such as in a visual search task.

A classical neural architecture proposed by Itti et al. [5] proposes a pure bottom-up approach to generate a saliency master map that is constructed in the posterior parietal cortex of primates (illustrated in Fig. 3). The feature integration theory (FIT) [15], which is a well-known cognitive theory, states that different features are processed in distinct brain regions. In line with the guidelines set out in the FIT [5], Itti's model utilizes color and intensity variations to produce several layered feature maps that depict the differences seen in prominent areas. A winner-take-all (WTA) decision model selects the most salient region out of a master map that linearly combines conspicuity maps in which different feature layers are fused. In addition, an inhibition of return (IOR) mechanism ensures that when transversing the visual field to complete a visual task, previously visited regions are avoided by the WTA mechanism.

The traditional Itti-Koch attention model [5] inspired multiple modern approaches to human visual cognition, and among those is VOCUS2 [6], an improved and modernized version of this classical model. This model comprises multiple necessary adaptations that maximize the performance of the model, with special emphasis on the scale-space, enabling a flexible center-surround ratio definition. VOCUS2 [6] is not only coherent in structure, but also fast and, as a typical bottom-up conspicuity model, performs saliency computations solely at the pixel level. The Invariant Visual Search Network (IVSN) [8] is an alternative model that takes advantage of the immense power of CNNs to extract top-down high-level features from a visual scene, in order to perform a visual target search. Similarly to the traditional attention model [5], the proponents of IVSN consider both WTA and IOR mechanisms when iteratively selecting the most promising regions from the field of view.

D. Active Perception and Attention Models

Active perception is a type of active learning [2] in which an agent gathers information from sensors and combines the current state of information with prior knowledge of the world to determine the next action to be taken. This approach can be used with various types of sensors and stimuli (for example, tactile, auditory, and olfactory) [2], yet our study focuses on the use of visual sensory information to perform search and exploration tasks. In the context of these visual tasks, the next move consists of a gaze shift toward a different focal point. Therefore, in this context, the active decision consists of selecting the most promising region to where the gaze should then be shifted. To guide the selection process, Figueiredo et al. [28] considered the usage of acquisition functions, such as the probability of improvement or the expected improvement, to choose the point that maximizes a task-specific function, through stochastic optimization. This is accomplished through the application of appropriately pre-defined metrics, such as the classification entropy or the categorical probability of a particular class.

Visual attention covers all factors that influence information selection mechanisms [3], whether guided by intrinsic aspects of the visual stimuli (bottom-up features) [5], [29] or by task-related expectations (top-down features) [8], [9]. On the one hand, the traditional saliency-based attention mechanism, proposed by Itti et al. [5], is the baseline model that inspired multiple modern approaches to human visual cognition, such as VOCUS2 [6], an improved and modernized version of this classical model, and low-level models that explore the relevance of contextual guidance [29] for eye movements and human attention. On the other hand, top-down models, such as the Invariant Visual Search Network (IVSN) [8] and SaltiNet [30], take advantage of the immense power of Convolutional Neural Networks (CNNs) to extract top-down high-level features, to perform visual search [15] in the presence of multiple distractors. Other advanced models take advantage of Deep Markov Models (ScanDMM [32]), spatial priority and scanpath networks (DeepGazeIII [33]), generative adversarial networks (PathGAN [31]), object detection models within a probabilistic framework [9], and visual transformers (HAT [34]) to model task-specific human visual scanpaths, using different types of visual inputs (e.g. Cartesian, foveal, 360-degree images) and suitable 2D/3D spatial representations [28], [35]. A significant number of new fixation prediction models are

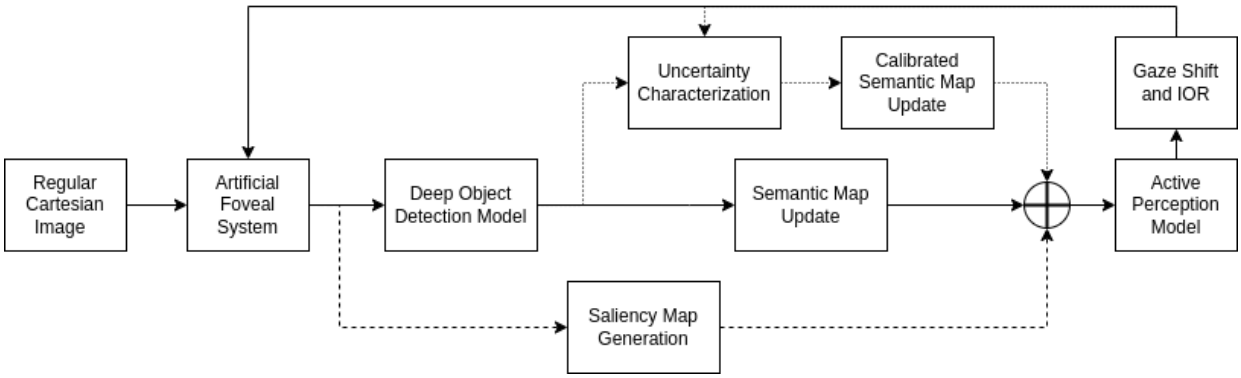

Fig. 4. Our general methodological approach (based on [9]) to semantic visual search and scene exploration tasks. An image is foveated on some initial image location, to simulate the human visual sensor. The foveal image is fed to an object detection model that may generate multiple bounding-boxes and the respective categorical scores. This information is used to update a world-fixed semantic map. This can be done in two ways: using the raw scores of the object detector (solid line) or with scores calibrated according to the effects of the foveal image characteristics (dotted line), since increased blur in the periphery will increase uncertainty to the classification scores. Alternatively, we also tested the use of a traditional saliency map approach to predict the next best focal point. The image is then foveated at the new selected focal point, simulating a saccade. A clasical Inhibition of Return method (IOR) [5] is applied to prevent returning to a previously visited location. The process is repeated until some terminal condition is met.

published every year. These models can be evaluated with properly maintained benchmarks, such as the MIT/Tuebingen Saliency Benchmark [36], which hosts specialized saliency datasets (COCO-Search18 [37], MIT300 [38], CAT2000 [39]) for visual tasks such as free-viewing and visual search, using well-established metrics [36]. Despite the considerable advancements that resulted from training modern deep learning models for visual attention [40], there is still a large margin for improvement when it comes to the integration of these models with biologically inspired mechanisms, derived from neurological and psychological principles and theories, such as hard attention [17], [19], [41], low-level bottom-up features [5], [6] and contextual information [29]. With this integration of concepts, through the combination of computer vision, visual cognition, and possibly Large Language models (LLMs) or Visual Language models (VLMs) that extract human knowledge on high-level relationships [42], emerges the opportunity to develop gaze fixation models that better mimic human visual-cognitive behaviors.

Recently developed models that consider either uniform Cartesian [7], [8] or foveal [17], [19], [21] visual systems, make use of CNNs to extract and process bottom-up, top-down, or both types of attentional features, which can be used to spot regions of interest over the entire visual field. Alternative approaches [9], [35] exploit the rich semantic information outputted by modern deep object detection models, building context grids that map the information collected from multiple object predictions distributed throughout the field of view. Some other approaches (e.g. [35], [43]) even consider modeling the full extent of humanoid behavior, allowing not only for eye movements but also full head and neck mobility.

III. Methodology

The fundamental concept supporting our methodology, illustrated in Fig. 4, is based on the idea that for most human activities, especially those involving visual tasks, it is highly beneficial to utilize previous sensory data [2] to plan for the next action. Such a strategy can efficiently minimize the number of actions (i.e. gaze shifts) necessary to complete the task at hand. Following the approach proposed by Dias et al. [9] we leverage the semantic information obtained by the state-of-the-art deep object detection models. This data is essential for pinpointing specific areas of importance within a scene, enabling the agent to make more informed decisions on where to focus its gaze next [2], particularly when engaging in human-like visual activities like visual search [15], and utilizing artificial foveal vision [20] to replicate the human field of vision.

A. Outline of the Model

1) Spatial Representation: In the context of robotic vision, there are multiple ways [28] to represent the geometry of the surrounding environment (e.g. voxel grids, point clouds). In our approach, we consider a fixed field of view, meaning that the agent can only perform simple ocular movements (i.e. saccades) without dynamically changing the configuration and the boundaries of the visual field. Taking this assumption into account, we select a typical two-dimensional Cartesian representation (namely, an occupancy grid [28]) that spatially encodes the surrounding environment into $\mathbf{x} = (x, y)$ coordinates. This representation is illustrated in Fig. 5 and is used to implement the *Semantic* and *Saliency* maps of Fig. 4. Moreover, in our approach, we consider and assume that the visual field has a static configuration [9], where the dispositions of the objects contained in it are immutable.

2) The Semantic Model: We assume that the agent can observe, at each spatial location (x, y), one object of possible K classes. The absence of objects is represented by the class $k = 0$ (background class). At a given time stamp t, with the gaze fixed on a certain region with $\mathbf{f}_t = (x_t, y_t)$ coordinates, the object detection model [4] outputs a set of object predictions $\mathcal{I}_t$, composed of L_t detections. As depicted

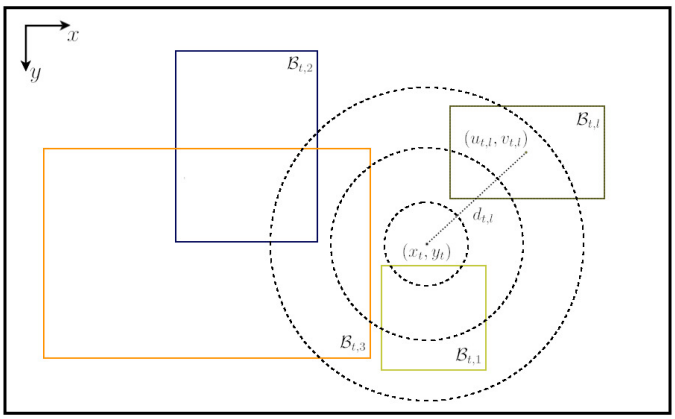

Fig. 5. Illustration of a foveal scene, fixed on a focal point, from which an arbitrary detector outputs L_t object predictions. The distance $d_{t,l}$ between the center of a bounding-box $\mathcal{B}_{t,l}$ and the center of the fovea is also represented.

in Fig. 5, each detection $I_{t,l} \in \mathcal{I}_t$, consists of a bounding-box $\mathcal{B}_{t,l}$ and a normalized score vector[1] $S_{t,l}$ of dimension $K+1$ (including the background class), that is,

$$S_{t,l} = (s_{t,l,0}, s_{t,l,1}, \ldots, s_{t,l,K}), \tag{1}$$

where $0 \leq s_{t,l,k} \leq 1, \forall k \in \mathcal{C}$ and $\sum_{i=0}^{K} s_{t,l,i} = 1$. $\mathcal{C} \subseteq [0, \ldots, K]$ is the set of possible classes and $l = 1, \ldots, L_t$. Through probabilistic inference, we want to create a map that aggregates all the observations obtained in the past ($\mathcal{I}_{1:t}$). The fused information should represent the semantic knowledge accumulated about the object classes considered $C^{\mathbf{x}} \in \mathcal{C}$, existing in the possible fixation points $\mathbf{x}$ in the context grid [10]

$$\mathbf{M}_t(\mathbf{x}) = [P(C^{\mathbf{x}}=0 \mid \mathcal{I}_{1:t}), \ldots, P(C^{\mathbf{x}}=K \mid \mathcal{I}_{1:t})]^T, \tag{2}$$

where $P(C^{\mathbf{x}} = k \mid \mathcal{I}_{1:t})$ represents the posterior probability of class $k \in \mathcal{C}$, given all past locally accumulated semantic information.

3) Semantic Map Update: The most straightforward way to update the semantic map (2) uses the raw classifier score vectors (1), as conveyed by the straight line path in Fig. 4. We use classifier fusion methods inspired in [13], which will be presented in Section III-B.

A more sophisticated way to update the semantic map, illustrated by the dotted line path in Fig. 4, takes advantage of the fact that observations at different eccentricities in the fovea yield different levels of uncertainty. Object detection models that have been pre-trained on a large set of regular Cartesian images are completely unaware of the existence of radial blur levels [1] that characterize foveal images [3]. This distortion imposed by a foveal visual system is responsible for a significant portion of the data-induced uncertainty observed in the generated detections. A fine-tuned calibrated probabilistic classifier of $(K+1)$ classes, similar to those incorporated within contemporary deep object detection models such as [14], can correctly quantify the level of uncertainty inherent in its instance-wise predictions. For this reason, we apply an efficient technique for uncertainty calibration [11], considering scores (1) as random variables sampled from a Dirichlet distribution $S \sim \mathrm{Dir}(\boldsymbol{\alpha})$ and use the Bayes theorem to find the class posterior distribution $P(C = k|S)$, as presented in Section III-C. In Section III-D we present a calibration method that accounts for the space-variant distortion that accompanies the prediction of each object.

4) Active Perception: Finally, in accordance with the block diagram presented in Figure 4, an active perception model extracts the information contained in one of the available maps to predict the next state of (2) and select the next best focal point, based on some pre-defined metric, to be presented in Section III-E. Following the cognitive model proposed by Itti et al. [5], as applied in modern visual models (e.g., VOCUS2 [6], IVSN [8]), we also consider an inhibition of return (IOR) mechanism that prevents the gaze from being directed to previously visited regions.

B. Fusion Model for Semantic Maps

In a Bayesian context, the probabilities $P(C^{\mathbf{x}}=k \mid \mathcal{I}_{1:t})$, which corresponds to the posterior distributions that make up the semantic map (2), can be treated as random variables with associated uncertainty. One simple and convenient solution to characterize the underlying uncertainty consists of modeling the distributions of the map (2) as Dirichlet distributions, $P(C^{\mathbf{x}}=k|\mathcal{I}_{1:t}) \sim \mathrm{Dir}(\boldsymbol{\beta})$, with parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)$. We can thus represent the semantic map (2) as a map of the Dirichlet β parameter estimates at each time and location

$$\mathbf{B}_t(\mathbf{x}) = \boldsymbol{\beta}_t^{\mathbf{x}} = \left[\beta_{t,0}^{\mathbf{x}}, \beta_{t,1}^{\mathbf{x}}, \ldots, \beta_{t,K}^{\mathbf{x}}\right]^T. \tag{3}$$

This allows for the inference of the posterior categorical probability distribution on different coordinates of the map, taking into account that $P(C^{\mathbf{x}}=k \mid \boldsymbol{\beta}_t^{\mathbf{x}})$ is a Dirichlet-compound multinomial distribution [12]

$$P(C^{\mathbf{x}}=k \mid \boldsymbol{\beta}_t^{\mathbf{x}}) = \frac{n!}{\left(\sum_{i=0}^{K} \beta_{t,i}^{\mathbf{x}}\right)^n} \prod_{i=0}^{K} \frac{\left(\beta_{t,i}^{\mathbf{x}}\right)^{n_i}}{n_i!} \tag{4}$$

where $n = \sum_{i=0}^{K} n_i, n_k = 1$ and $n_i = 0, \forall i \in \mathcal{C} \setminus k$. From these parameters we can compute the expected values and covariances of the class distribution:

$$p_k = E[C^{\mathbf{x}} = k] = \frac{\beta_k}{\sum_j \beta_j} \tag{5}$$

$$\sigma_{ij} = Cov(C^{\mathbf{x}} = i, C^{\mathbf{x}} = j) = -p_i p_j, i \neq j \tag{6}$$

In a visual task, we typically do not have access to the true semantic label of the observed objects, but only to the vector of scores provided by the object detector. The work developed by Kaplan et al. in [13] proposes some methodologies to fuse the information coming from probabilistic sensors. It establishes a comparison between Bayesian approaches (e.g. product rule,

[1]In case the detector's output (1) is not normalized, the scores can be normalized by dividing each $s_{t,l,k}$ by $\sum_{i=0}^{K} s_{t,l,i}$.

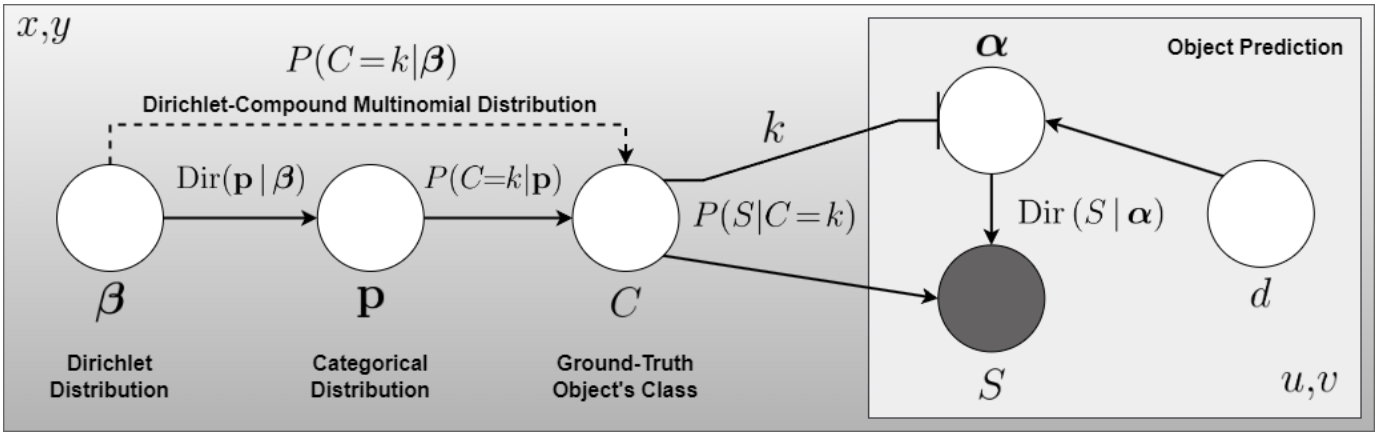

Fig. 6. Representation of the dependencies between the variables that participate in the methodology, as proposed by Dias et al. [9], in the form of a Bayesian network, using plate notation, explaining both semantic content mapping and foveal score calibration techniques. Essentially, we attach a Dirichlet probability distribution, with parameters $\boldsymbol{\beta}$, to each pair of coordinates (x, y) of the context grid, which, in the context of Bayesian inference, serves as the prior for a categorical distribution $\mathbf{p}$ that represents the confidences in the presence of each class $k \in \mathcal{C}$. Furthermore, we simplify the notation by defining the actual posterior distribution for the ground truth class C through a Dirichlet compound multinomial distribution [12], as detailed in Section III-B.

sum rule) and a proposed fusion rule (to which Dias et al. [9] conveniently refer to as Kaplan rule), formulated with (3) as

$$\beta_{t+1,k} = \frac{\beta_{t,k}\left(1 + \dfrac{\lambda_{t,l,k}}{\sum_{j=0}^{K}\beta_{t,j}\lambda_{t,l,j}}\right)}{1 + \dfrac{\min_j \lambda_{t,l,j}}{\sum_{j=0}^{K}\beta_{t,j}\lambda_{t,l,j}}}. \tag{7}$$

where $\lambda_{t,l,k} = P(S_{t,l}|C = k)$ are the detector score likelihoods. Hence, assuming that our detector is perfect and that, for this reason, we do not have to account for epistemic (or model) uncertainty [14], we can simply consider the generated detection scores (1) as the likelihoods $P(S_{t,l}|C = k) = s_{t,l,k}$. Therefore, in this particular case, we consider $s_{t,l,k} = \lambda_{t,l,k}$.

Kaplan's rule (7) is a moment-matching approach [13] that can be applied when fusing new measurements (1) retrieved from surrounding location detections. One clear advantage of this fusion rule [13] is that, as a moment-matching approach, it can return a posterior Dirichlet distribution that fits the actual class posterior distribution (4) in terms of the first statistical moment (i.e. the mean). Essentially, the expected value of the resulting Dirichlet distribution (3) tends to be an accurate estimate [13] of the actual posterior distribution's mean (2).

C. Score Calibration

Kaplan's rule (7) uses the detector score likelihoods instead of the actual scores (see [13] Eq. (1)), defined in Fig. 6 as $P(S|C = k)$. If a detector is calibrated, then $s_k = P(C = k|S) \propto \lambda_k$ [14], and the actual classifier scores can be used in Kaplan's rule (7). However, object classifiers are often not properly calibrated to reflect the posterior probabilities of the actual class, contrary to what is often assumed in many classification problems [44]. To obtain the actual measurement likelihoods $P(S|C = k)$ we learn, from a large training dataset, multiple Dirichlet distributions for each class $\mathrm{Dir}(S \mid \boldsymbol{\alpha}_k)$, which can be applied in a Bayes classifier

$$P(C = k|S) = \frac{P(S|C = k)\,P(C = k)}{P(S)} \propto \mathrm{Dir}\,(S|\boldsymbol{\alpha}_k)\,, \tag{8}$$

considering $P(C = k)$ as a uniform prior categorical distribution, thus avoiding the introduction of bias in the model. The Dirichlet likelihoods for each class are estimated through

$$\mathrm{Dir}(S \mid \boldsymbol{\alpha}_k) = \frac{\Gamma\left(\sum_{i=0}^{K}\alpha_{k,i}\right)}{\prod_{i=0}^{K}\Gamma(\alpha_{k,i})}\prod_{i=0}^{K} s_i^{\alpha_{k,i}-1} \tag{9}$$

involving the Euler's Gamma function $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$.

Essentially, each set of Dirichlet parameters is estimated by fitting multiple detection data, consisting of score vectors (1), respectively, associated with a ground-truth class $k \in \mathcal{C}$. There is no known closed-form maximum-likelihood estimation for Dirichlet distributions. Hence, we consider a simple and efficient iterative method, proposed by Minka [44], to fit the semantic data extracted from a large dataset to the Dirichlet likelihood (9). This culminates in several sets of α parameters

$$\boldsymbol{\alpha}_k = (\alpha_{k,0}, \alpha_{k,1}, \ldots, \alpha_{k,K}) \tag{10}$$

with which we capture the aleatoric uncertainty, derived from multiple data-related factors, such as illumination or occlusion.

D. Foveal Observation Model

The incorporation of a foveal visual system generates extra data-induced uncertainty in the outputs of the object detection model [3], as shown in Fig. 7. Taking advantage of the radial distribution of blur across the visual field, a foveal observation model [9] is designed to calibrate the scores (1), directly outputted by the detector. However, the calibration model now depends not only on the object class but also on the distance to the fovea, as the classifier scores are significantly affected by the different levels of blur across the simulated visual field.

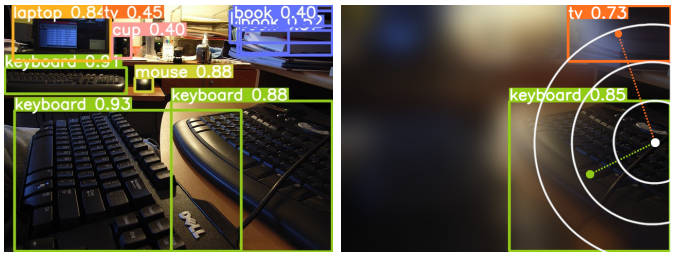

Fig. 7. Example of an ambiguous detection due to peripheral distortion [1]. The semantic content located on the top-right corner of the simulated visual field corresponds to the class *'book'* (left). Nonetheless, the class predicted by the detector is *'tv'* (right), due to the influence of the foveal sensor's position.

This calibration technique considers the relative distance from the center of the prediction's bounding-box $\mathcal{B}_{t,l}$ to the center of the fovea, $d = \|(u, v)\|$, as illustrated in Fig. 5. The pair (u, v) represents the set of central coordinates of $\mathcal{B}_{t,l}$ relative to the center of the fovea. The foveal observation model, introduced in Fig. 4 as the method used for uncertainty characterization and represented in Fig. 6 by a plate, essentially comprises multiple sets of Dirichlet parameters for each class and distance to the fovea (in a discrete set), specifically a structure that contains $(K+1) \times N$ arrays of $\boldsymbol{\alpha}_{k,d}$ parameters. This means that there is a set of parameters (10) for each class in all N distance levels. These levels consist of discretizations of the focal distance (d) that grant the existence of a finite number of compartmentalized Dirichlet distributions that are trained to model the $p\,(S|C = k, d)$ likelihood distribution. From this likelihood, we compute $p\,(C = k|S, d)$, with a small adaptation of the Bayes Rule (8). In this adaptation, we consider the $\mathrm{Dir}(S \mid \boldsymbol{\alpha}_{k,d})$ distribution as the estimated likelihood, $p\,(S|C = k, d)$, with which we can model the $p\,(C = k|S, d)$ posterior distribution. Hence, a new set of calibrated scores $S'_{t,l}$ proportional to the measurement likelihoods can be obtained through (8)

$$\begin{aligned} S' &= (s'_0, s'_1, \ldots, s'_K) \\ &= \frac{1}{D}\left[p(S|C=0,d), \ldots, p(S|C=K,d)\right]^T \\ &= \frac{1}{D}\left[\mathrm{Dir}(S \mid \boldsymbol{\alpha}_{0,d}), \ldots, \mathrm{Dir}(S \mid \boldsymbol{\alpha}_{K,d})\right]^T \end{aligned} \tag{11}$$

where $D = \sum_{k=0}^{K} \mathrm{Dir}(S \mid \boldsymbol{\alpha}_{k,d})$ represents the normalization factor which grants that $\sum_{k=0}^{K} s'_k = 1$. Notice that, since Kaplan's rule (7) uses directly the likelihoods instead of the posterior probabilities, the normalization of the likelihoods is not ultimately relevant for this process, and is hereby defined only to demonstrate that the calibrated scores (11) can be conformed into a categorical distribution. The prior distribution $p(C = k \mid d)$, which consequently appears in the adaptation of (8), is assumed to be uniform, as an object can be arbitrarily located within the confines of a scene. Kaplan's classifier update rule (7) can also accommodate foveal calibrated scores since we can directly replace raw scores (1) with corrected foveal scores (11) in (7). This is identified in Fig. 4 as the calibrated semantic map, which can be found along the dotted path, right after the uncertainty characterization block.

E. Active Perception Model

Let us now turn our focus to the active perception model, which operates based on the information collected on the semantic map (3). At a given timestamp t, the current state of (3) is consulted by the active perception mechanism, to determine the most promising cell (x^*_{t+1}, y^*_{t+1}) where the gaze is to be shifted at the $t+1$ instant, according to a pre-defined task-specific metric.

1) Non-Predictive Visual Search: Taking into account the dominant greedy nature of the visual search task [15], the priority is to find the region where the target object is localized and to orient the gaze toward it. Therefore, we start by considering a simple non-predictive optimization problem, consisting of maximizing the posterior (2) of a given target class $k \in \mathcal{C}$, as

$$\left(x^*_{t+1}, y^*_{t+1}\right) = \mathbf{x}^* = \underset{\mathbf{x}}{\mathrm{argmax}} \left\{ \frac{\beta^{\mathbf{x}}_{t,k}}{\sum_{j=0}^{K} \beta^{\mathbf{x}}_{t,j}} \right\}. \tag{12}$$

This intuitive approach consists of consulting the current state of the map (3) in order to infer the class confidence in each pair of coordinates (x, y) in the context grid. Then, similarly to traditional cognitive approaches [5], [6], we use a winner-take-all (WTA) mechanism that selects the next best focal point.

2) Predictive Visual Search: Although it is not the first attempt to combine object detection and active perception to perform visual search [35], this is the first work to integrate the semantic-based active perception model, introduced by Dias [9], into a visual target search task with foveal vision. Therefore, we wish to evaluate the full potential of the active perception approach, by also including an update simulation, in order to predict the posterior probabilities (2). Taking this objective into account and following the proposed methodology [9], we simulate a complete map update (3) for each possible focal cell, depending on $\mathbf{f}' = (x', y')$, where the fovea could be next centered. We define the predicted semantic maps as

$$\bar{\mathbf{B}}^{\mathbf{f}'}_{t+1}(\mathbf{x}) = \mathbb{E}\left[\mathbf{B}_{t+1}(\mathbf{x}) \,|\, \mathbf{B}_t(\mathbf{x}), \mathbf{f}'\right], \tag{13}$$

which simulates the expected value of the semantic map after a fixation to coordinate $\mathbf{f}'$ is executed. This means that the active perception mechanism has to select one of $X \times Y$ possible scenarios, based on the knowledge that results from each simulation. To solve (13), we first compute the expected value of the scores that should be outputted by the object detector for the coordinates $\mathbf{x}$, if the focal point is set in $\mathbf{f}'$,

through the expansion of the posterior distribution

$$\begin{aligned} & p\left(S^{\mathbf{x}} \mid \mathbf{f}', \boldsymbol{\beta}_t^{\mathbf{x}}\right) = \\ & = \sum_{k=0}^{K} p\left(S^{\mathbf{x}}, C^{\mathbf{x}}{=}k \mid \mathbf{f}', \boldsymbol{\beta}_t^{\mathbf{x}}\right) \\ & = \sum_{k=0}^{K} p\left(S^{\mathbf{x}} \mid C^{\mathbf{x}}{=}k, \mathbf{f}', \boldsymbol{\beta}_t^{\mathbf{x}}\right) P(C^{\mathbf{x}}{=}k \mid \mathbf{f}', \boldsymbol{\beta}_t^{\mathbf{x}}) \\ & = \sum_{k=0}^{K} p\left(S^{\mathbf{x}} \mid C^{\mathbf{x}}{=}k, \mathbf{f}'\right) P(C^{\mathbf{x}}{=}k \mid \boldsymbol{\beta}_t^{\mathbf{x}}) \,. \end{aligned} \tag{14}$$

The Dirichlet-compound multinomial [12] posterior class probabilities $P\left(C^{\mathbf{x}} = k \mid \boldsymbol{\beta}_t^{\mathbf{x}}\right) = \beta_{t,k}^{\mathbf{x}} \,/\, \sum_{j=0}^{K} \beta_{t,j}^{\mathbf{x}}$ can be inferred through (4). The steps presented in (14) take advantage of the conditional independence between $S^{\mathbf{x}}$ and $\boldsymbol{\beta}_t^{\mathbf{x}}$ for a given class k. It also depends on the fact that the class to which the object contained in a $\mathbf{x} = (x, y)$ cell belongs is not constrained by the new location of the center of the fovea $\mathbf{f}' = (x', y')$. Taking advantage of the linearity of the expectation operator applied to (14), we can then assert that

$$\mathbb{E}\left[S^{\mathbf{x}} \mid \mathbf{f}', \boldsymbol{\beta}_t^{\mathbf{x}}\right] = \sum_{k=0}^{K} \mathbb{E}\left[S^{\mathbf{x}} \mid C^{\mathbf{x}}{=}k, \mathbf{f}'\right] P(C^{\mathbf{x}}{=}k \mid \boldsymbol{\beta}_t^{\mathbf{x}}) \tag{15}$$

meaning that the computation of predictive scores is hanging on the determination of $\mathbb{E}\left[S^{\mathbf{x}} \mid C^{\mathbf{x}}{=}k, \mathbf{f}'\right]$ in each iteration. Consider now that $d' = \|\mathbf{x} - \mathbf{f}'\| = \|(u', v')\|$ represents the distance level associated with a given $\mathbf{x}$ cell in relation to the central coordinates of the $\mathbf{f}'$ cell, where the gaze will be shifted, measured with the $(u', v') = (x - x', y - y')$ relative coordinates, where (x, y) and (x', y') are the central pixel coordinates of the referred $\mathbf{x}$ and $\mathbf{f}'$ cells, respectively. Hence, the expected score vectors can be computed as

$$\bar{S}^{\mathbf{x}} = \mathbb{E}[S^{\mathbf{x}} \mid \mathbf{f}', \boldsymbol{\beta}_t^{\mathbf{x}}] = \sum_{k=0}^{K} \frac{\boldsymbol{\alpha}_{k,d'}}{\sum_{j=0}^{K} \alpha_{k,d',j}} \frac{\beta_{t,k}^{\mathbf{x}}}{\sum_{j=0}^{K} \beta_{t,j}^{\mathbf{x}}} \tag{16}$$

which can then be applied with Kaplan's rule (7) to update the predicted semantic map (13). From then on, we can infer the estimated posterior distributions (2) out of $\bar{\mathbf{B}}_{t+1}^{\mathbf{f}'}$ (13), through (4), determining the cell most promising, to where the gaze is to be shifted, based on a given task-dependent metric [28].

For the predictive visual search task, given its greedy nature [15], we will not need complete updates of the predicted semantic maps. In fact, only the scores of the detections in the high-resolution area close to the next fixation point $\mathbf{f}'$ are relevant. In other words, we assume that the best fixation point will be close to the object of interest. Thus, for each possible fixation point $\mathbf{f}'$, we only need to predict the detector scores at the foveal locations ($d = 0$)

$$\bar{S}_0^{\mathbf{x}} = \sum_{k=0}^{K} \frac{\boldsymbol{\alpha}_{k,0}}{\sum_{j=0}^{K} \alpha_{k,0,j}} \frac{\beta_{t,k}^{\mathbf{x}}}{\sum_{j=0}^{K} \beta_{t,j}^{\mathbf{x}}}, \tag{17}$$

using only the set of Dirichlet distributions, with parameters (10), that correspond to the innermost distance level.

Similarly to the non-predictive approach (12), we use a WTA mechanism that selects the next best focal point through

$$\left(x_{t+1}^*, y_{t+1}^*\right) = \mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmax}} \left\{ \frac{\bar{\beta}_{t+1,C}^{\mathbf{x}}}{\sum_{k=0}^{K} \bar{\beta}_{t+1,k}^{\mathbf{x}}} \right\}, \tag{18}$$

where $\bar{\beta}_{t+1,C}^{\mathbf{x}}$ represents the expected Dirichlet parameter for class C, extracted from the local $\mathbf{f}' = \mathbf{x}$ predictive map $\bar{\mathbf{B}}_{t+1}^{\mathbf{f}'=\mathbf{x}}$.

3) Scene Exploration: In the exploration task, the goal is to minimize the overall uncertainty in the map while performing a minimal number of actions (i.e. gaze shifts). The gaze shift will be made to the new location of the fixation that minimizes the uncertainty of the predicted map (13). One form of quantifying uncertainty consists of computing the distance between the semantic map distributions and non-informative distributions [9]. We consider the Kullback-Leibler divergence [45] as a measurement of the distance between the estimated Dirichlet distribution and a base distribution, as

$$F_t^{\mathbf{x}} = \mathcal{D}_{KL}\left(\operatorname{Dir}\left(\boldsymbol{\beta}_t^{\mathbf{x}}\right) || \operatorname{Dir}\left(\boldsymbol{\beta}^0\right)\right), \tag{19}$$

where $\boldsymbol{\beta}^0$ corresponds to the initial uniform Dirichlet distribution ($\beta_k^0 = 0.5\,, \forall k \in \{0, \ldots, K\}$), representing a state of maximum uncertainty, that is initially warranted to each $\mathbf{x}$ cell. Intuitively, we want to select the cell that maximizes the distance between the two Dirichlet distributions [9] with (19). The classification negentropy [45] is another uncertainty metric considered for active perception, formulated as

$$F_t^{\mathbf{x}} = \sum_{k=0}^{K} \frac{\beta_{t,k}^{\mathbf{x}}}{\sum_{j=0}^{K} \beta_{t,j}^{\mathbf{x}}} \log\left(\frac{\beta_{t,k}^{\mathbf{x}}}{\sum_{j=0}^{K} \beta_{t,j}^{\mathbf{x}}}\right). \tag{20}$$

The higher the negentropy, the lower the cell's confusion. For this reason, we consider applying both the Kullback-Leibler divergence (19) and the negentropy (20) metrics with a simple acquisition function that transverses each map (13), accumulating the expected results of the selected metric. With this approach, we select the location $\mathbf{f}' = (x', y')$ that maximizes the sum of the expected metric results over (14) as

$$\left(x_{t+1}^*, y_{t+1}^*\right) = \mathbf{x}^* = \underset{\mathbf{f}'}{\operatorname{argmax}} \left\{ \sum_{\mathbf{x}} \mathbb{E}\left[F_{t+1}^{\mathbf{x}} \mid \mathbf{f}'\right] \right\}, \tag{21}$$

where $\mathbb{E}\left[F_{t+1}^{\mathbf{x}} \mid \mathbf{f}'\right]$ represents the expected measurement of the metric F for the next iteration (i.e. at the $t+1$ timestamp), assuming that the focal point is then set on the $\mathbf{f}'$ coordinates. Finally, the last alternative metric consists of computing the absolute difference between the categorical scores of the two most probable classes (2), which can be expressed with (3) as

$$F_t^{\mathbf{x}} = \max_k \left\{ \frac{\beta_{t,k}^{\mathbf{x}}}{\sum_{j=0}^{K} \beta_{t,j}^{\mathbf{x}}} \right\} - \max_{k \setminus \varkappa} \left\{ \frac{\beta_{t,k}^{\mathbf{x}}}{\sum_{j=0}^{K} \beta_{t,j}^{\mathbf{x}}} \right\}, \tag{22}$$

where $\varkappa = \underset{k}{\operatorname{argmax}} \left\{ \beta_{t,k}^{\mathbf{x}} \,/\, \sum_{j=0}^{K} \beta_{t,j}^{\mathbf{x}} \right\}$ represents the most probable class, taking advantage of the knowledge in (3). Following the work of Dias et al. [9], we consider applying this metric with a different acquisition function [28], known as

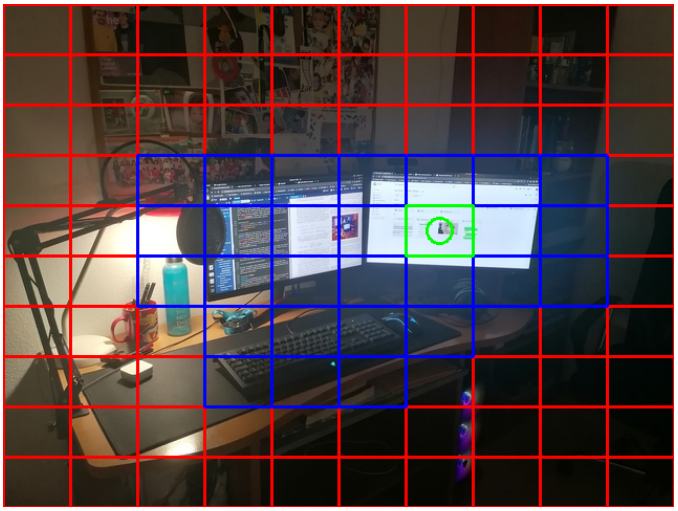

Fig. 8. Grid of (red) cells dividing the semantic map of Fig. 1. The green-colored cell that is visible on the 10x10 grid marks the focal point and the cells that are updated with the YOLOv3 detections are highlighted in blue.

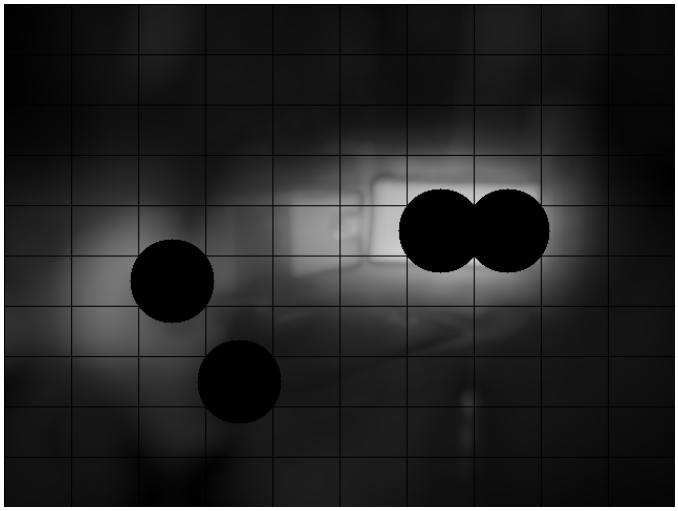

Fig. 9. The saliency map generated by VOCUS2 [6] for the same foveal scene, after 4 algorithm iterations (applying the IOR mechanism), is also presented.

the expected improvement. This function aims at maximizing the expected magnitude of the two-peaks improvement (22) as

$$\left(x^*_{t+1}, y^*_{t+1}\right) = \mathbf{x}^* = \underset{\mathbf{f}'}{\operatorname{argmax}} \left\{ \max_{\mathbf{x}} \left| \mathbb{E}\left[F^{\mathbf{x}}_{t+1} | \, \mathbf{f}'\right] - F^{\mathbf{x}}_t \right| \right\} \quad (23)$$

which is a greedy approach, since it consists of selecting the maximum absolute difference, out of all $\mathbf{x}$ map coordinates, for each $\mathbf{f}'$, that represents all the possible next focal points.

IV. Experiments & Results

In this section, we intend to highlight some important aspects of our testing setup and present the most relevant experimental results, consisting of the evaluation of cognitive attention models, proposed and presented in Section III, during the completion of active visual perception experiments involving scene exploration and visual search tasks.

A. Overview of the Experimental Setup

First, to generate predictions, as depicted in Fig. 5, we select YOLOv3 [25], which is a deep detection model that follows the one-step framework [4], based on global regression/classification. Although not the latest version, YOLOv3 remains a fast and accurate model for object detection that uses Darknet-53 [25] as its CNN backbone architecture.

The foveal observation model is trained with the full COCO 2017 training set [46], foveated around random focal points, using Minka's algorithm [44] to fit the categorical scores (1) generated by YOLOv3, into multiple Dirichlet distributions (10), according to the corresponding ground truth object and focal distance level. Our validation dataset comprises 300 quasi-randomly selected images from the COCO validation set. Essentially, each selected image must contain at least a target object and multiple distractors, adding up to a grand total of 8 or more different instances. Furthermore, any target object contained in a selected image can only occupy and be associated with, at most, 20% of the total area of the fixed field of view, to promote a fair iterative search process and assess the influence of peripheral vision. The image is divided into a grid of 10×10 to implement the semantic map cells, as depicted in Fig. 8, which delimits the regions to which the gaze can be shifted. When initializing an experiment, each cell of the semantic map (3) is defined by a uniform Dirichlet distribution $\boldsymbol{\beta}^0$, where $\beta^0_k = 0.5, \forall k \in \{0, \ldots, K\}$, representing an initial non-informative state. The initial focal point is also pseudo-randomly selected, as the only constraint being applied consists of not selecting a starting cell containing an instance of the targeted class. Suppose that multiple detections overlap a cell with (x, y) coordinates. In that case, Kaplan's rule (7) is applied sequentially, in no particular order, with every single score vector, with (11) or without (1) foveal calibration.

Regarding scene exploration, our aim is to conclude whether the active perception model is able to reduce the confusion inherent to the semantic map (3) with the minimal amount of gaze fizations. For this reason, we assess the performance of the methods, presented in Section III-E, with the success rate metric, which consists of the ratio between the number of map cells where the highest categorical score (4) fits the class of an overlapping ground-truth object (true positives), and the total number of cells containing ground-truth objects (true positives + false negatives).

Regarding the visual search task, we figure out whether it is possible to navigate a visual scene, seeking an instance of a pre-defined class, when applying the approach described in Fig. 4. Despite the broad acceptance criteria for success in visual search, related to the number of target cells, it is crucial to understand to what extent the semantic information model is able to leverage the accuracy with which the task is completed. Following the setup presented in [8], a *Oracle* is responsible for terminating the search algorithm once the focal point is set on a target cell, indicating that the search has been concluded because one instance of the target class has been found. We evaluate the success in each iteration as the ratio between the

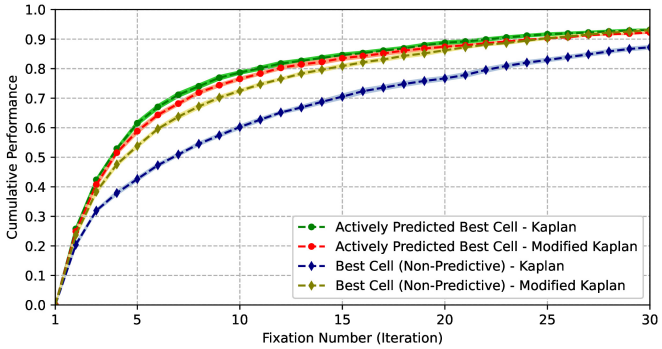

Fig. 10. Comparison between the mean values of the cumulative performance and the respective standard error of the mean (SEM) bands from results observed over 10 experiments, obtained using predictive and non-predictive approaches that operate on the semantic information updated either with raw (Kaplan) or calibrated scores (Modified Kaplan).

number of images in which the target object has already been located at that iteration and the total number of images in the validation dataset, i.e. the cumulative performance [8].

B. Baseline Saliency-Based Model

To attempt a fair comparison with an attention-based model that imitates biological cognitive mechanisms, we consider a classical conspicuity model [5] that is adapted to govern perceptive decisions (dashed line in Fig. 4). This constitutes what is commonly called a free-viewing model [10] as the decision process is not influenced by any task-related intricacies.
For this purpose, we implement VOCUS2 [6], a reformed version of Itti's traditional saliency model [5], to extract low-level features inherent to pixel-level color contrasts, in compliance with the FIT. Despite lacking top-down features (considered in recent visual attention models [8]), which are essential for approximating human neurological models to their full potential, VOCUS2 is still able to simulate critical aspects that guide the mechanisms that govern human perception. In Fig. 9 we present a visualization of the IOR mechanism, applied on the saliency map produced by VOCUS2 after each iteration, nullifying the intensity of the pixels contained inside the previously visited regions (confined in a grid of cells [9]).

C. Visual Search

In the search experiments, we assess how many gaze shifts are necessary to achieve success, with each experiment running for a total of 30 iterations. For each image from the validation dataset, we repeat the experiment 10 times, starting from different initial focal points (i.e. cells) to test whether the model is exposed to different information in every repetition. Success is achieved when one of the evaluated models leads the gaze to be shifted toward a cell that contains an instance of the class that is being targeted [15] in each experiment.

1) Role of Prediction and Calibration: We compare the two proposed methods of active visual search: non-predictive active perception (12), using only the current information gathered and stored on the semantic map (3), and predictive active perception (18), with simulated updates. Furthermore, we investigate potential benefits that may proceed from implementing the foveal correction model (11), determining whether the calibrated map can facilitate the efficient identification of cells that contain instances of the target class. Two main conclusions arise from the results that are portrayed in Fig. 10. The first conclusion emerges from the comparison between the non-predictive search algorithms. From the results highlighted in Tab. I we can conclude that, using the corrected scores (11), calibrated by the foveal observation model III-D, appears to be considerably beneficial for the non-predictive approach, but not influencing the predictive approach. Foveal calibration leads the non-predictive methods to a performance level that approximates the results obtained with the predictive methods while consuming much fewer computational resources. The second conclusion is that the active predictive approaches [9] comfortably outperform the non-predictive approaches, regardless of the use of the foveal calibration mechanism presented in Section III-D. Therefore, predictive approaches (18) can provide added accuracy, at the cost of additional computation.

TABLE I
COMPARISON OF PREDICTIVE AND NON-PREDICTIVE APPROACHES

Evaluation Metric	Predictive		Non-Predictive	
	Non-Calib	***Calib***	**Non-Calib**	***Calib***
CP[1] @ 5	61,53%	58,83%	42,63%	53,80%
CP[1] @ 15	84,60%	83,53%	70,53%	80,90%
CP[1] @ 30	93,10%	92,23%	87,23%	93,03%
Max. SEM[2]	0,772%	0,820%	0,845%	0,718%
Comp. Cost[3]	77,81 ms	83,76 ms	0,144 ms	0,133 ms

[1]Cumulative Performance. [2]Standard Error of the Mean. [3]Per Iteration.

2) Comparison with Saliency-Based and Random Search: In a second experiment with visual search methods, we compare the proposed active approach (with foveal calibrated scores) with random and saliency-based approaches. Although the most relevant comparison would be with the performance achieved by human subjects, this work aims to make the comparison with a conspicuity model [5] inspired by the behavior of the early primate visual system. However, previous studies (e.g. [8], [35]) have shown that humans tend to generally outperform unguided visual search algorithms. The results are displayed in Fig. 12 and Tab. II. From the mean values of the cumulative performance, it is possible to conclude that the semantics-based methods perform better than the unguided random search approach in the initial iterations (up to 10^{th} in the non-predictive mode and up to the 13^{th} in the predictive mode). We also noticed that the non-predictive approach with non-calibrated scores initially takes the upper hand, yet eventually falls in the face of the random search algorithm. Regarding the performance of VOCUS2 in object search, it is quite plausible that the lack of precision when finding target objects is directly related to the fact that the search algorithm does not consider class-specific information. Therefore, the simulated attention mechanism, built on bottom-up saliency information that is extracted from the stimuli at the

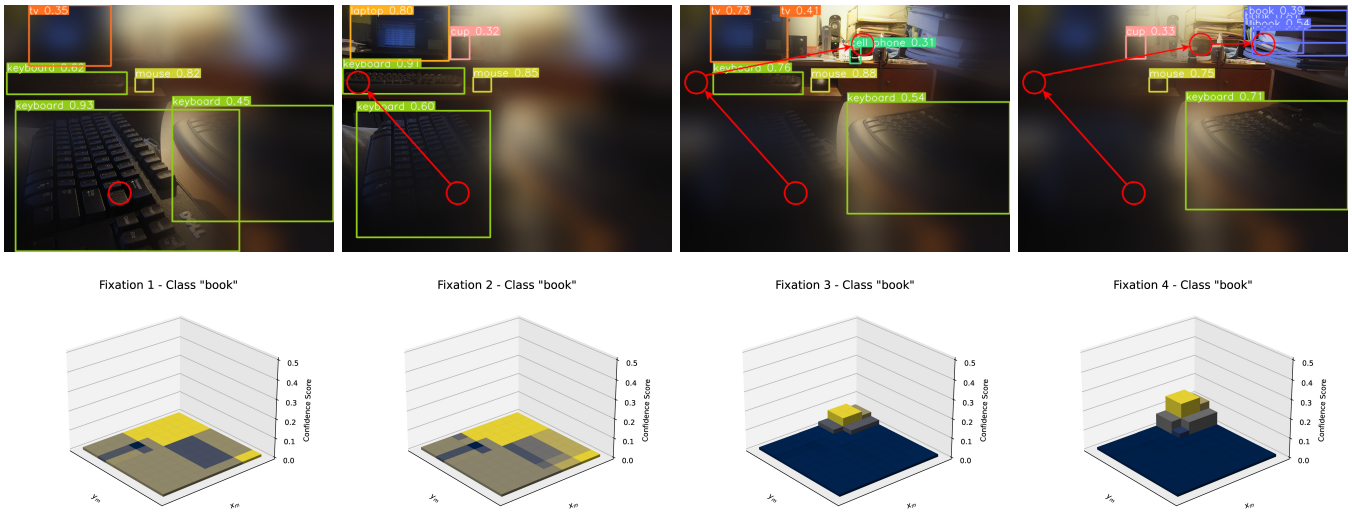

Fig. 11. Example of a visual search experiment, using the full semantic-based methodology, presented in Section III, with foveal uncertainty characterization. The active perception approach applied in this example is fully predictive (18), selecting the cell that maximizes the class posterior probability (2) after simulating a complete map update. The target class, pre-defined for this experiment, is *'book'* and there are multiple instances of this class in the top-right corner of the image that is being used to simulate a visual field. Here we present the YOLOv3 [25] object predictions generated at each fixation and the resulting saccade paths (figures in the top), as well as the target class confidence scores (figures at the bottom), for each cell of the map, in the form of bar graphical plots. After only 4 fixations, the semantic-based active perception model is able to direct the gaze toward the region where the target object is actually located.

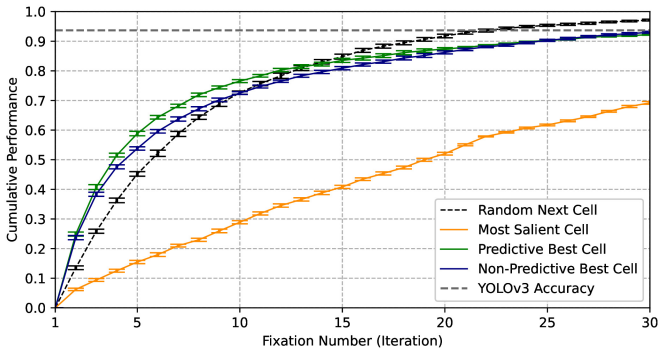

Fig. 12. Comparison between the mean values of the cumulative performance and the respective SEM bars, obtained using semantic and saliency-based [6] approaches, together with the results of the random search. Both predictive and non-predictive models exploit the foveal calibration model.

TABLE II
COMPARISON OF ACTIVE VISUAL SEARCH APPROACHES

Eval. Metric	***Random***	***Saliency***	***Predictive***	***Non-Predictive***
CP[1] @ 5	45,20%	15,47%	58,83%	53,80%
CP[1] @ 15	84,80%	41,82%	84,60%	80,90%
CP[1] @ 30	97,13%	69,13%	93,10%	93,03%
Max. SEM[2]	1,059%	0,627%	0,820%	0,718%
Comp. Cost[3]	0,015 ms	762,3 ms	83,76 ms	0,133 ms

[1]Cumulative Performance. [2]Standard Error of the Mean. [3]Per Iteration.

pixel level, appears to be very inefficient in finding objects in a class-oriented context, in the presence of multiple distractors.

3) An Illustrative Sample Case: Let us consider the visual search example, presented in Fig. 11. The target class is *'book'* and the results illustrate the application of the predictive approach (18) (Section III-E) with the foveal calibration mechanism (Section III-D). During the execution of the task, the confidence score of the target class p_C (6) increases consistently, fixation after fixation, in the cells that surround and overlap the location of the target ground truth objects [10], while progressively decreasing in all other cells of (3). In this example, after only four algorithm iterations (corresponding to 3 complete saccades), the agent would be able to effectively direct its attention toward the region where objects of the category (class) *'book'* are factually located. Furthermore, notice how YOLOv3 [25] is not able to directly detect the targeted instances until the last iteration of the algorithm. In this example, the initial focal point was purposely set on a region that is close to the target's opposite corner location, in order to observe the influence of the initial focal point, avoiding direct visibility of the targeted objects at the beginning of the experiment. The observed results align with a fundamental aspect of AET [15], which states that, in the context of visual search, the ability to detect and distinguish between target and distractor objects is quite crucial to the human cognitive system. In this example, we observe that the detection of distractors leads the model to actively neglect the regions where these types of objects are contained. This is related to the fact that during the update process (7) (Section III-B) lower target class scores (1) tend to decrease the influence of the parameter of (3) associated with that same class, in relation to the parameters associated with higher class scores. Moreover, due to this distinction of target-distractor dissimilarities, the

model tends to favor unexplored regions, where the semantic content is yet unknown. The target's initial confidence bar plot, which can be observed in Fig. 11 (yellow cells), shows that the cells that are not overlapped by distractor-related detections present higher target class posteriors (2), in relation to the cells where these types of object are detected. Since in this example we apply the predictive approach (18), the confidences p_k (6) represented in the bar plots do not reflect the estimated confidences that are considered by the active perception model. Hence, to obtain the state (3) of the estimated map that is fed to the optimization function (18), one must first simulate an update for each cell, using Kaplan's rule (7), with the estimated local scores $\bar{S}_0^{\mathbf{x}}$ (17), to obtain $\bar{p}_{t+1,k}$ (6) as described in Section III-E. Therefore, the most promising cell, selected by the active perception mechanism, to where the gaze is shifted, does not correspond to the cell associated with the highest target class probability (2), based on the current state of the semantic map (3). Finally, observing the evolution of the target's categorical probability (2) at different gaze fixations, we conclude that the confidence scores observed in the first iterations are hardly distinguishable, as the absolute differences between them are almost negligible. For the latter fixations, the states of the map clearly differentiate the region where the ground-truth target objects are located from the rest of the image. The target's class confidences start to distance themselves from the confidences associated with the cells that do not cover that same region, therefore validating the proposed methodology.

D. Scene Exploration

Regarding exploration experiments, our objective is to capitalize on previously performed experiments [9] to verify how accurately the knowledge collected on the maps represents the semantic content displayed within the confines of the visual scene. Furthermore, we assess the relative computational costs of the different metrics, applied to their respective acquisition functions [28], as presented in Section III-E3, to compare the performance of this model with the performances achieved by both saliency-based [6] and random gaze selection approaches when it comes to accurately represent the semantic content that is available in the scene.

First, we performed a comparative analysis of the different acquisition functions, on the map obtained using foveal calibrated scores (11). The mean and SEM of the success rates in each saccade, obtained for all images from the validation set, are presented in Fig. 13. From the results displayed in Fig. 13 we elect the maximization of the Kullback-Leibler divergence as the best-performing active perception metric among those considered. This metric (19) consistently outperforms negentropy (20) and, in relation to the two-peak difference gain (22), presents a better trade-off between the amount of correct information collected in the first iterations and the long-term performance.

Secondly, we compare the proposed semantic model [9], using the Kullback-Leibler divergence metric (19), with a random algorithm, which randomly selects the next gaze direction, and VOCUS2 [6], the elected saliency-based method. Furthermore, we assess the use of calibrated (11) and non-calibrated (1) scores in the map update stage. Fig. 14 presents the average success rates obtained. By maximizing the Kullback-Leibler divergence, semantic exploration not only outperforms the saliency-based active exploration but also initially leverages the amount of semantic content that is correctly mapped even in the face of a random selection. Moreover, there is no significant upside in considering the calibrated scores rather than the raw class scores. After about 4 saccades, the random gaze selection algorithm eventually outperforms the other active perception methods. The reason for this phenomenon is that, after a certain amount of iterations, the fovea has already been placed over all the critical regions of the images, directly capturing the majority of the

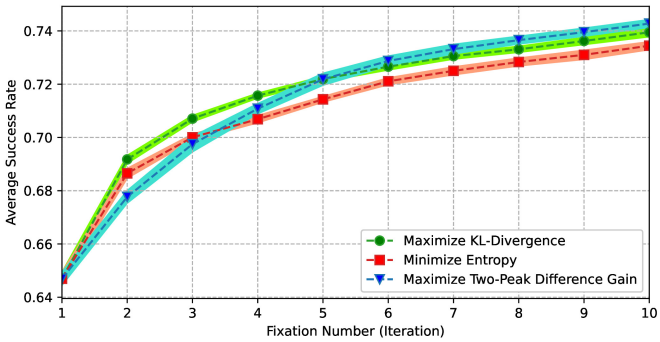

Fig. 13. Comparison of the mean values of the (average) success rate obtained after 10 repetitions, each starting in a different focal point, with different active perception techniques, together with their respective standard error bands.

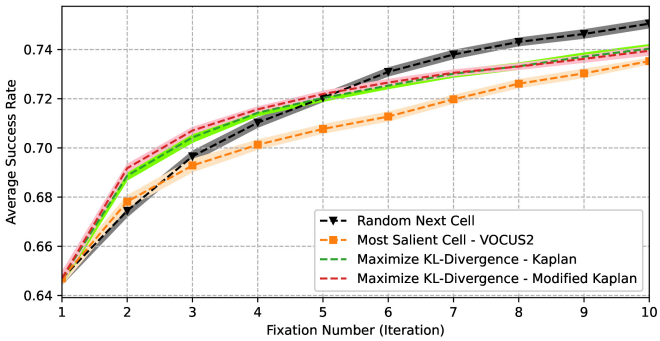

Fig. 14. Mean values of the average success rate and the respective standard error (SEM) bands, for the semantic-based active perception approach that considers the probability of improving (21) the KL-Divergence metric (19) using either the raw (Kaplan) or calibrates (Modified Kaplan) scores, in comparison with both the random and VOCUS2 [6] results.

TABLE III
COMPUTATIONAL COSTS[1] OF SCENE EXPLORATION APPROACHES

Random	***Saliency***	***KL-Divergence***	***Entropy***	***Two-Peak***
0,017 ms	0,462 s	9.851 s (2)	9.770 s (2)	8.971 s (2)
		9,907 s (3)	9,833 s (3)	8,981 s (3)

[1]Per Iteration. Updates: [2]w/ Kaplan rule. [3]w/ Modified Kaplan rule.

semantic content that is possible to extract with YOLOv3 [25] or any other object detection model.

Finally, from the data presented in Tab. III, we draw the conclusion that our implementation of the semantic-based model for exploration [9], described in Section III, implies a larger computational cost than the saliency-based model. The disparity between the time-cost values obtained with VOCUS2 [6] and the proposed methodology [9] comes from the overhead inflicted by the quadratic complexity of the map update process on the exploration algorithm. The fact that for each individual cell, out of the $X \times Y$ grid, we must compute a total of $X \times Y$ updates, simulating a gaze shift toward each (x', y') cell, is reflected in a computational cost about $20\times$ higher compared to the cost of generating a saliency map. However, the proposed methodology increases the amount of accurately mapped information in the first few saccades.

V. Conclusion

This work has studied the performance of the semantic model [9] in visual search and exploration tasks, using different active perception metrics, and performing a qualitative and quantitative comparison with VOCUS2 [6], a bottom-up saliency model. We have fully implemented and tested the active perception methodology proposed by Dias et al. [9], adapting it to the visual search context [15]. To this end, we designed and meticulously followed the methodological pipeline illustrated in Fig. 4, fully testing the saliency-based approaches [6] and the semantic-based approaches, with and without foveal calibration [11], as described in Section III. By using object detection models pre-trained with Cartesian images, it is much faster to fit calibrate the classifier scores with a [44]-based foveal observation mechanism within a Bayesian framework [10], than to retrain the detection model with a foveal image dataset.

A. Highlighted Contributions

Let us start by considering the results obtained during the experimental phase with regard to the scene exploration experiments. First, we have been able to show, through extensive experimentation, that the methodology proposed by Dias et al. [9] is able to beat the performance of a biologically inspired attention mechanism [5]. Moreover, we have validated the ability of the different active perception techniques presented in Section III-E3, when it comes to accurately mapping the available semantic content of the scene. We also wished to understand whether the use of calibrated scores (11), outperforms the use of direct raw scores (1) of YOLOv3 [25]. The results of Section IV-D have shown that there is no particular advantage in performing the calibration.

Regarding visual search, the foveal observation model, which calibrates the scores that are used to update a map (3), has a positive impact on the performance of the semantic model. Furthermore, it is also established that the predictive approach [9], using the expected value of the probability distribution in each cell to simulate an update, is undoubtedly advantageous, beating the traditional bottom-up saliency model [6] by a substantial margin. Notice that VOCUS2 performs significantly under other visual search methods, due to the information considered not being target-oriented. Furthermore, in terms of computational cost, from Tab. II we infer that, compared to the cost of generating a full conspicuity map, our proposed active perception mechanism for visual search (18) is not only effective, but also quite fast.

Visual search results support the idea that the human cognitive recognition system does not rely solely on bottom-up features [15], but also on top-down features and other target similarity principles [47] grounded in the AET. The lack of top-down features in the model [6] justifies the inaccuracy exhibited, when completing the search task, in contrast to other recently developed models [8] that incorporate such features.

B. Future Work

Following the consideration made in Section III-E regarding the greedy nature of the visual search task, it is important not to neglect the semantic information available in the cells surrounding a potential fixation point, which can be extracted from (13). It is possible that, by incorporating information available in adjacent cells, both the accuracy and precision of the active perception model increase. However, to achieve spatial precision, weights should be attributed to the information contained in the surrounding cells, according to their distance to the focal point. For future consideration, it would certainly be interesting to investigate whether the integration of information available in adjacent cells could positively influence the performance of the semantic-based active perception model [9]. Such a technique would get more out of the semantic map, taking full advantage of the spatial properties of the cells and the relationships between them, at the cost of slightly increasing the overhead of the algorithm.

Moreover, it would also be interesting to compare the semantic-based model considered in this work [9], with an adapted version of IVSN [8], which can iteratively search for stimuli with foveal vision, extracting top-down characteristics along the way. Furthermore, this comparison could also be extended to include novel visual search approaches that are based on deep learning models (e.g. ScanDMM [32], PathGAN [31], DeepGaze III [33], HAT [34], or even a model based on a Bayesian ideal observer [48]) that have delivered better results when it comes to information gain between consecutive saccades and scan path similarity with human subjects. Understand how Large Language models (LLMs) and Visual Language models (VLMs) can provide human knowledge [42] that can be exploited to make both search and exploration more efficient.

Finally, it would be interesting to adapt the proposed methodology, to perform the same humanoid task, yet with a real mobile agent that must be able to perform body movements [35], moving through the environment while dynamically changing the range, location, and direction of the field of view.

REFERENCES

[1] E. E. Stewart, M. Valsecchi, and A. C. Schütz, "A review of interactions between peripheral and foveal vision," Journal of Vision, vol. 20, no. 12, pp. 2–2, 2020.

[2] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," Autonomous Robots, vol. 42, pp. 177–196, 2018.

[3] R. P. de Figueiredo and A. Bernardino, "An overview of space-variant and active vision mechanisms for resource-constrained human inspired robotic vision," Autonomous Robots, pp. 1–17, 2023

[4] Z. Q. Zhao, P. Zheng, S. -T. Xu and X. Wu, "Object Detection With Deep Learning: A Review," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212-3232, Nov. 2019.

[5] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[6] S. Frintrop, T. Werner, and G. Martin Garcia, "Traditional saliency reloaded: A good old model in new shape," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 82–90, 2015.

[7] R. Burt, N. N. Thigpen, A. Keil, and J. C. Principe, "Unsupervised foveal vision neural architecture with top-down attention," Neural Networks, vol. 141, pp. 145–159, 2021.

[8] M. Zhang, J. Feng, K. T. Ma, J. H. Lim, Q. Zhao, and G. Kreiman, "Finding any Waldo with zero-shot invariant and efficient visual search," Nature Communications, vol. 9, no. 1, p. 3730, 2018.

[9] A. Dias, L. Simões, P. Moreno and A. Bernardino, "Active Gaze Control for Foveal Scene Exploration," 2022 IEEE International Conference on Development and Learning (ICDL), London, United Kingdom, pp. 115-120, 2022.

[10] M. Kümmerer and M. Bethge, "Predicting visual fixations," Annual Review of Vision Science, vol. 9, pp. 269–291, 2023.

[11] M. Xie, S. Li, R. Zhang, and C. H. Liu, "Dirichlet-based Uncertainty Calibration for Active Domain Adaptation," The Eleventh International Conference on Learning Representations, 2023.

[12] N. L. Johnson, S. Kotz, and N. Balakrishnan, Discrete multivariate distributions, vol. 165. Wiley New York, 1997.

[13] L. M. Kaplan, S. Chakraborty and C. Bisdikian, "Fusion of classifiers: A subjective logic perspective," 2012 IEEE Aerospace Conference, Big Sky, MT, USA, pp. 1-13, 2012.

[14] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," Machine Learning, vol. 112, no. 9, pp. 3211–3260, Sep. 2023.

[15] L. K. Chan and W. G. Hayward, "Visual search," Wiley Interdisciplinary Reviews: Cognitive Science, vol. 4, no. 4, pp. 415–429, 2013.

[16] V. J. Traver and A. Bernardino, "A review of log-polar imaging for visual perception in robotics," Robotics and Autonomous Systems, vol. 58, no. 4, pp. 378–398, 2010.

[17] P. Ozimek, N. Hristozova, L. Balog, and J. P. Siebert, "A space-variant visual pathway model for data efficient deep learning," Frontiers in Cellular Neuroscience, vol. 13, p. 36, 2019.

[18] D. Pamplona and A. Bernardino, "Smooth Foveal vision with Gaussian receptive fields," 2009 9th IEEE-RAS International Conference on Humanoid Robots, Paris, France, pp. 223-229, 2009.

[19] H. Lukanov, P. König, and G. Pipa, "Biologically inspired deep learning model for efficient foveal-peripheral vision," Frontiers in Computational Neuroscience, vol. 15, p. 746204, 2021.

[20] A. F. Almeida, R. Figueiredo, A. Bernardino, and J. Santos-Victor, "Deep networks for human visual attention: A hybrid model using foveal vision," in ROBOT 2017: Third Iberian Robotics Conference: Volume 2, pp. 117–128, 2018.

[21] C. Melício, R. Figueiredo, A. F. Almeida, A. Bernardino and J. Santos-Victor, "Object detection and localization with Artificial Foveal Visual Attention," 2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), Tokyo, Japan, pp. 101-106, 2018.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587, 2014.

[23] W. Liu et al., "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37, 2016.

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016.

[25] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," University of Washington, 2018.

[26] Y. Fang et al., "Eva: Exploring the limits of masked visual representation learning at scale," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19358–19369, 2023.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European conference on computer vision, 2020, pp. 213–229.

[28] R. P. de Figueiredo, A. Bernardino, J. Santos-Victor, and H. Araújo, "On the advantages of foveal mechanisms for active stereo systems in visual search tasks," Autonomous Robots, vol. 42, pp. 459–476, 2018.

[29] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.," Psychological review, vol. 113, no. 4, p. 766, 2006.

[30] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Saltinet: Scan-path prediction on 360 degree images using saliency volumes," in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2331–2338, 2017.

[31] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "PathGAN: Visual scanpath prediction with generative adversarial networks," in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0, 2018.

[32] X. Sui, Y. Fang, H. Zhu, S. Wang, and Z. Wang, "ScanDMM: A deep markov model of scanpath prediction for 360deg images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6989–6999, 2023.

[33] M. Kümmerer, M. Bethge, and T. S. Wallis, "DeepGaze III: Modeling free-viewing human scanpaths with deep learning," Journal of Vision, vol. 22, no. 5, pp. 7–7, 2022.

[34] Z. Yang, S. Mondal, S. Ahn, G. Zelinsky, M. Hoai, and D. Samaras, "Predicting Human Attention using Computational Attention," 2023.

[35] R. Druon, Y. Yoshiyasu, A. Kanezaki and A. Watt, "Visual Object Search by Learning Spatial Context," in IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 1279-1286, April 2020.

[36] M. Kummerer, T. S. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 770–787, 2018.

[37] Y. Chen, Z. Yang, S. Ahn, D. Samaras, M. Hoai, and G. Zelinsky, "Coco-search18 fixation dataset for predicting goal-directed attention control," Scientific reports, vol. 11, no. 1, p. 8776, 2021.

[38] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.

[39] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," 2015.

[40] M. Kümmerer and M. Bethge, "State-of-the-art in human scanpath prediction," 2021.

[41] G. Elsayed, S. Kornblith, and Q. V. Le, "Saccader: Improving accuracy of hard attention models for vision," Advances in Neural Information Processing Systems, vol. 32, 2019.

[42] R. Fu, J. Liu, X. Chen, Y. Nie, and W. Xiong, "Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning," arXiv preprint arXiv:2403.11401, 2024.

[43] M. Grotz, T. Habra, R. Ronsse and T. Asfour, "Autonomous view selection and gaze stabilization for humanoid robots," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, pp. 1427-1434, 2017.

[44] T. Minka, "Estimating a Dirichlet distribution," Massachusetts Institute of Technology, Tech Report, 2000.

[45] C. M. Bishop and N. M. Nasrabadi, Pattern recognition and machine learning, vol. 4. Springer, 2006.

[46] T.Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Computer Vision – ECCV 2014, pp. 740–755, 2014.

[47] J. M. Wolfe, "Visual search: How do we find what we are looking for?," Annual review of vision science, vol. 6, pp. 539–562, 2020.

[48] S. Rashidi, W. Xu, D. Lin, A. Turpin, L. Kulik, and K. Ehinger, “An active foveated gaze prediction algorithm based on a Bayesian ideal observer,” Pattern Recognition, vol. 143, p. 109694, 2023.