Cost-Efficient Multi-Scale Fovea for Semantic-Based Visual Search Attention

João Luzio
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal
joaoluzio14@tecnico.ulisboa.pt

Alexandre Bernardino
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal
alex@isr.tecnico.ulisboa.pt

Plinio Moreno
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal
plinio@isr.tecnico.ulisboa.pt

***Abstract*—Semantics are one of the primary sources of top-down preattentive information. Modern deep object detectors excel at extracting such valuable semantic cues from complex visual scenes. However, the size of the visual input to be processed by these detectors can become a bottleneck, particularly in terms of time costs, affecting an artificial attention system's biological plausibility and real-time deployability. Inspired by classical exponential density roll-off topologies, we apply a new artificial foveation module to our novel attention prediction pipeline: the Semantic-based Bayesian Attention (*SemBA*) framework. We aim at reducing detection-related computational costs without compromising visual task accuracy, thereby making *SemBA* more biologically plausible. The proposed multi-scale pyramidal field-of-view retains maximum acuity at an innermost level, around a focal point, while gradually increasing distortion for outer levels to mimic peripheral uncertainty via downsampling. In this work we evaluate the performance of our novel *Multi-Scale Fovea*, incorporated into *SemBA*, on target-present visual search. We also compare it against other artificial foveal systems, and conduct ablation studies with different deep object detection models to assess the impact of the new topology in terms of computational costs. We experimentally demonstrate that including the new *Multi-Scale Fovea* module effectively reduces inherent processing costs while improving *SemBA*'s scanpath prediction accuracy. Remarkably, we show that *SemBA* closely approximates human consistency while retaining the actual human fovea's proportions.**

***Index Terms*—Foveal Vision, Human Attention, Visual Search, Object Detection, Scanpath Prediction**

I. Introduction

Human attention is guided by multiple sources of preattentive information [1], such as top-down and bottom-up features, task-related rewards, prior knowledge, scene context, and semantics. Information coming from all these sources is then combined into a single spatial priority map [2], [3], often referred to as the attention map. The relevance of each source is conditioned by the task at hand. On the one hand, bottom-up salient cues and scene syntax may be of utmost relevance for exploratory tasks [4], such as free-viewing. On the other hand, top-down semantics and task-related rewards play a crucial role in goal-directed activities [5], such as visual search.

This work was supported by *Fundação para a Ciência e Tecnologia* (FCT) through project (1801P.01460.1.03) LARSYS/ISR BASE 2025-2029 - VISLAB LAB/ISR (DOI: 10.54499/UID/50009/2025).
João Luzio is supported by the FCT doctoral grant [2024.00683.BD].

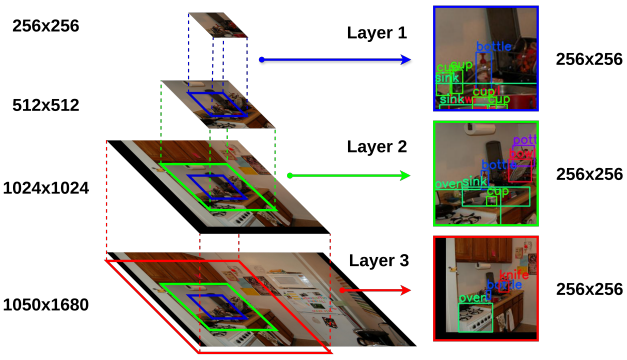


Fig. 1. Illustration of our novel *Multi-Scale Fovea* mechanism. The proposed method consists of building a multi-resolution pyramid [12], around a selected focal point, and then downsampling all levels to the size of the innermost layer, to mimic the eccentricity effect [8]. Object detections from outer levels tend to reflect the uncertainty that derives from such exponential pixel density reduction [13]. This technique facilitates the sequential extraction of semantic information [9] in a more cost-efficient and biologically plausible manner.

The extraction of information from these sources is also conditioned by the physiology of the human retina. The eye exhibits a higher photoreceptor density around a central focal region [6] known as the fovea. Within this region, objects are perceived with maximum visual acuity [7] and are therefore easier to identify. As the distance from the fixated point increases, photoreceptor density decreases, leading to progressively blurrier percepts in more peripheral regions. This phenomenon, known as the eccentricity effect [8], makes objects located further in the retinal periphery increasingly difficult to recognize, due to the intensifying distortion levels.

In essence, the behavior of the human active perception mechanism [2] hinges on the nature of the visual task, the characteristics of the field-of-view, and the content of the scene. In this work, we address attention prediction in a target-present visual search setting, where the goal is to find a target object that is known to be present on a provided visual *stimulus*. Mainstream attention models operate by localizing and collecting preattentive information [5] to determine the most salient regions and infer plausible human-generated scanpaths. By scanpaths, we refer to sequences of fixation points [4], [28].

In recent years, multiple human scanpath prediction models have been proposed: e.g. Gazeformer [32], HAT [33],

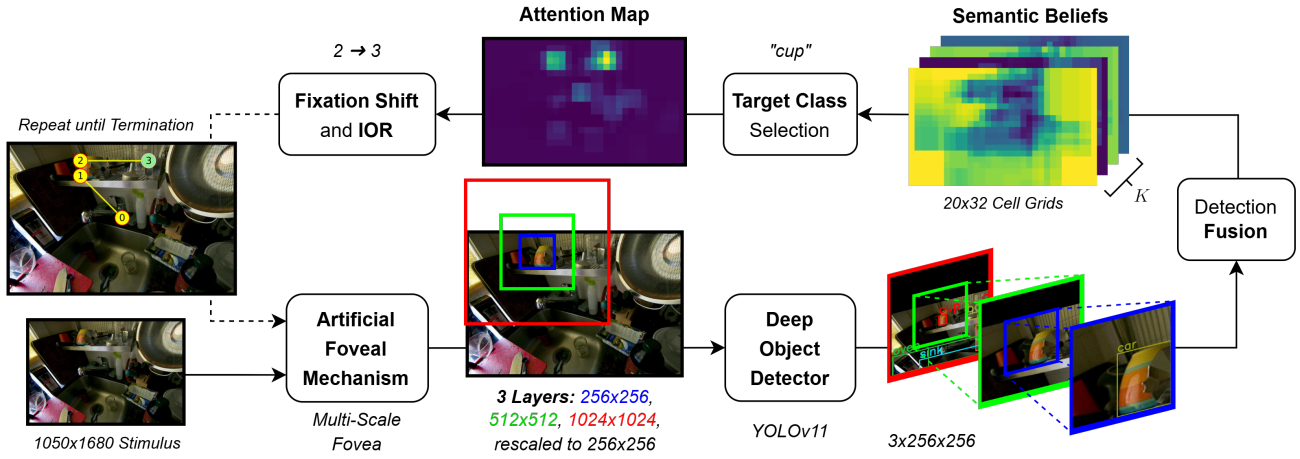


Fig. 2. Schematic representation of the Semantic-based Bayesian Attention framework, i.e. *SemBA* [10], for attention prediction, applied to visual search. In this example, we apply the proposed *Multi-Scale Fovea* mechanism, that generates $N = 3$ layers around a focal point, starting at a base dimension of 256×256 pixels and doubling in size between consecutive layers. All layers are then down-sampled to the base dimension to emulate peripheral distortion. Note that the artificial foveation block can be replaced by any other mechanism, such as *FOVEA* [17] or *Laplacian Foveation* [14]. The same follows for the deep object detection block, where although we apply YOLOv11 [23] any other modern detection model can be integrated, e.g. DETR [24], RT-DETR [25]. Object detections are then filtered and fused on multiple grids of 20×32 cells, to build semantic belief maps for all K known classes. *SemBA*'s active perception system then selects the map for a specific targeted class, which is *cup* in this example, and selects the next-best fixation point based on its current belief state. This process is repeated until a termination criterion is met, while applying inhibition of return (IOR [27]) to prevent revisiting fixated locations.

CLIPgaze [34]. Most of these methodologies exploit selective attention mechanisms, present in modern deep learning architectures, to perform feature extraction in a foveated manner. Despite deriving from classical attention methods [2], [3], state-of-the-art models [32]–[34] have begun to incorporate additional data modalities into their learning process and are now trained using eye-tracking data collected from human subjects. However, though dependent on top-down and bottom-up features, our built-in attention system for visual search [1] does not rely on external data provided by other human beings. For instance, infants and children don't learn to visually explore their surrounding environment or search for specific toys from scanpaths generated by their caregivers. In this sense, we can assert that visual attention is purely *stimulus* driven [5]. As an alternative to this notable deviation from biologically-inspired attention modeling, we have recently proposed *SemBA* [10], a semantic-based probabilistic framework for human scanpath prediction. This purely *stimulus*-driven attention prediction pipeline, illustrated in Fig. 2, leverages pre-trained deep object detectors (e.g. YOLO [21], DETR [24]) to extract and fuse top-down semantic cues from multiple fixations. *SemBA* has been shown to be able to perform target-present visual search [9], competing with other state-of-the-art models in terms of human scanpath similarity. However, it is known that input dimensionality constitutes a bottleneck for time and memory costs in object detection [18], hindering *SemBA*'s performance. Nevertheless, due to foveal eccentricity [8], our cognitive system cannot process an entire scene at once. Instead, it cumulatively integrates foveal and peripheral knowledge [7] gathered across multiple eye saccades and fixated regions.

In summary, the human foveal system restricts both the amount and the quality of information that can be perceived from a certain viewpoint. This conditioning is reflected on how objects are displayed across different regions of the field-of-view in terms of geometrical appearance and visual acuity. Most modern deep object detection models are pre-trained on large conventional Cartesian image datasets, such as the COCO 2017 dataset [26], being able to invariantly detect objects at different resolutions. To exploit this geometric property of detectors we recover the classical foveal visual systems [12] and propose a *Multi-Scale Fovea* mechanism (Fig. 1) for localized semantic content extraction. Our new foveation module builds a multi-level pyramid by cropping concentric regions of increasing size around a fixation point. All levels are then downsampled to the scale of the innermost layer in order to increasingly degrade the resolution of semantic content, located in the periphery, while preserving each level's geometrical configuration. On the one hand, our *Multi-Scale Fovea* avoids the need to retrain or fine-tune detectors to handle alternative fovea-like input topologies [14], [16], [17]. On the other hand, it promotes an effective computational cost reduction in object detection [13], making semantic-based attention frameworks, e.g. *SemBA* [10], more suitable for deployment in bio-inspired real-time systems, such as humanoid robots. Regarding our work, we highlight the following contributions:

- We propose a novel *Multi-Scale Fovea* mechanism for cost-efficient semantic-based visual attention modelling.
- We incorporate the new fovea module into *SemBA* [10], and assess its performance in target-present visual search.
- We assess human-model scanpath similarity, using off-the-shelf metrics [28], and compare *SemBA*'s scores and efficiency with state-of-the-art model's results [31]–[34].

Code availability: An implementation of *SemBA* × *Multi-Scale Fovea* is available at github.com/vislab-tecnico-lisboa/SemBA.

II. Background and Related Work

A. Deep Object Detection

Object detection is a famous computer vision problem which involves jointly localizing objects via bounding-boxes and assigning them semantic categorical labels. Early methods, largely based on hand-crafted features and multi-stage pipelines, suffered from limited robustness and scalability [18].

Capitalizing on the hierarchical feature extraction capabilities of convolutional neural networks (CNNs), the YOLO framework [21] was introduced as a one-stage family of deep detectors designed for real-time object recognition. YOLO approaches detection by treating it as a unified prediction task, where bounding-boxes and class labels are directly inferred in one forward pass. By leveraging global image context, earlier YOLO versions achieved faster inference while sacrificing some localization accuracy, especially for small or crowded objects. The YOLO family has evolved over the years [22] with new versions (e.g. YOLOv11 [23]) introducing refinements and enhancements to increase the model's performance.

The remarkable success of transformers in natural language processing [19] has prompted growing interest in extending their use to computer vision tasks. The detection transformer (DETR [24]) is considered the main foundational model for object detection using visual transformers. Whereas YOLO mainly relies on local feature processing, exploiting CNNs' convolutional kernels, DETR is able to model spatial and contextual relationships between multiple objects across an entire scene. However, despite its competitive performance, DETR is slow to converge during the training phase and shows limited effectiveness in detecting small objects. As a means to reduce computational costs and increase accuracy, recent iterations to the original DETR architecture have been proposed [19], such as DINO, Co-DETR, LW-DETR, and RT-DETR. In particular, the real-time detection transformer architecture (RT-DETR [25]) introduces adaptable inference speed, balancing high accuracy with real-time performance.

B. Computational Visual Attention Models

Human gaze control has been a topic of interest in neurology and psychology [4], given its importance for understanding visual perception and cognition. In recent years, this particular topic has also been gaining prominence among the computer vision community. The seminal work of L. Itti and C. Koch [3] paved the way for saliency-based attention, offering an intuitive and neurologically-inspired way to relate bottom-up features (i.e. color contrasts, intensity and orientation) to free-viewing attentional behavior. The Itti-Koch model was the first to exploit the concept of saliency maps, believed to be an integral part of the posterior parietal cortex of early primates.

Despite the advancements on bottom-up saliency models [28], goal-directed attention was mostly discarded until the rise of artificial neural networks in popularity. The invariant visual search network (IVSN [31]) was a pioneer work on top-down target-present visual search attention, specifically for zero-shot settings. This template matching model processes a generic target image in parallel with the full scene, using two distinct neural networks. An attention map is built by matching the outputs of both networks. To generate scanpaths, likely fixation points are then sequentially selected as the most conspicuous regions. However, the IVSN does not apply any human-like vision system, uniformly processing the entire scene at once.

Similar to object detection, computational attention has developed alongside deep learning [28], leveraging new architectures and mechanisms to more accurately predict scanpaths. Modern methods such as Gazeformer [32] and HAT [33] exploit the selective attention mechanisms present in vision transformers to predict sequences of fixation points for any provided image. Another model, known as CLIPgaze [34], leverages a foundational visual language model (CLIP) pre-trained with large amounts of textual and visual data to predict human-like scanpaths in zero-shot settings. While HAT integrates visual information at two different eccentricities, using peripheral and foveal tokens [33], Gazeformer and CLIPgaze do not explicitly account for the characteristics of the human field-of-view. These models have been able to achieve state-of-the-art performance, thriving on the establishment of the first visual search benchmark dataset: COCO-Search18 [27].

COCO-Search18 provides not only the scenes themselves but also the scanpaths of real human subjects for each instance of the dataset. Modern methods [32]–[34] are trained using the visual *stimuli* and their respective human-generated scanpaths. However, as mentioned in Section I, humans don't learn from other humans' fixation data. Our innate attention system is purely guided by information that can only be perceived directly from the field-of-view. Nevertheless, goal-directed attention relies on visual cues that are top-down by nature, such as semantic information. For this reason, we recently proposed a semantic-based Bayesian attention framework (*SemBA* [10]) for human attention prediction, which leverages pre-trained deep object detectors. In essence, *SemBA* functions as a pipeline for collecting and fusing semantic information (into an attention map) across multiple fixation points. The full implementation of *SemBA* is detailed in Section III-B.

C. Artificial Foveation Methods

In the literature, the definition of fovea is often ambiguous, resulting in different assumptions on its actual dimensions. Although the region of maximum acuity, known as foveola [6], comprises only a visual angle of 1°, intermediate eccentricities between the fovea and the periphery, often referred to as parafovea [7], can extend roughly from 2° to 5° in diameter.

To artificially reproduce foveated fields-of-view with such characteristics, computational models [14], [16], [17] apply different types of geometric transformations to regular images. Cesar Bandera and Peter D. Scott's pioneering work on foveal machine visual systems [12] set the stage for current research on bio-inspired vision. As an answer to the increasing pixel flow-through rate problem in image processing, their work proposed the usage of linear and exponential pixel density roll-offs in rectangular and hexagonal sampling lattices. Related work [13] proved that such technique allows for significant

time and memory cost reductions in object detection while maintaining critical geometrical features of Cartesian images.

Moving away from multi-resolution Cartesian methods and toward more biologically plausible vision, Almeida et al. [14] proposed a Laplacian pyramid that applies radial Gaussian filtering to remove high spatial frequencies on the periphery. This method, which from now on we refer to as *Laplacian Foveation*, has been demonstrated to not interfere with object detection's performance, provided that the non-blurred central region, corresponding to the fovea, covers a large enough area. *Laplacian Foveation* [14] has been already successfully incorporated into the *SemBA* framework [10] and applied on target-present visual search [9] putting on a convincing performance. However, this foveation technique does not promote any effective decrease in terms of computational costs, as object detectors still need to process images in full resolution.

As an attempt to mimic retinal sampling, recent vision models [15], [16] apply log-polar transformations on regular images. These more biologically-inspired geometric transforms not only effectively reduce the number of pixels to be processed but are also invariant to centered scaling and rotations. However, accurate object detection requires models to be retrained on large datasets of log-polar-mapped images, which consumes a substantial amount of time and computational resources. Another recent model, known as *FOVEA* [17], uses kernel density estimation from bounding-boxes, generated by an object detector, to construct saliency maps that guide grid-based image warping. High detection density regions are magnified through transforming the original scene into a new warped space, where pixels are sampled according to the estimated saliency. New detections generated in a warped space can then be remapped to the original unwarped scene.

III. Methodology

In this section, we first outline the preliminary geometric constraints [28] commonly employed in standard human attention prediction setups. We then introduce the probabilistic framework [9], [10] adopted for semantic data fusion and next-best fixation prediction in visual search. Finally, we describe our proposed fovea-inspired mechanism for efficient object detection, which reduces computational costs while introducing uncertainty that depends on the visual field's dynamics.

A. Scanpath Prediction Setup

In line with other state-of-the-art methodologies for human scanpath prediction [31]–[33], we treat two-dimensional images [27] as fixed fields-of-view, where the spatial resolution and scene boundaries remain static. We also assume a non-dynamical scene configuration [10], where the arrangement of objects within the landscape does not change over time.

Consider an image with dimensions *height* $\times$ *width* and an initial fixation point f_0, typically positioned at the center of the visual field. When searching or exploring a given visual *stimulus*, a human produces a sequence of fixation points $f_1, f_2, \ldots, f_T$, where each fixation $f_t, \forall t \in \{1, \ldots, T\}$, corresponds to a specific pixel location within the image. It should be emphasized that the length of the fixation sequence, T, differs across tasks, scenes, and individual subjects [27].

To reduce the action space, we employ a standard two-dimensional Cartesian representation to spatially encode the surrounding environment, structured as a $Y \times X$ context grid. By adopting this representation, we minimize computational complexity while approximating the cognitive system's sensitivity levels, balancing efficiency with biological plausibility.

B. Semantic-based Bayesian Attention Framework

To assess the impact of different foveal geometries in attention prediction, we must first select an appropriate pipeline for collecting and fusing information. These data, extracted across a sequence of fixation points, are used to cumulatively build attention maps from which next fixation locations are inferred. Hence, we consider our recent framework for semantic-based Bayesian attention, i.e. *SemBA* [10], that has already been successfully applied to human scanpath prediction [9] in a target-present visual search setting. This method exploits the rich semantic information extracted by modern deep object detectors [18] to perform goal-directed fixation prediction.

Start by considering a set $\mathcal{C} \subseteq \{1, \ldots, K\}$ containing all the classes recognizable by a given object detection model. An object detection is composed of a score vector $S = (s_1, s_2, \ldots, s_K)$, where $s_k \geq 0, \forall k \in \mathcal{C}$, and its respective bounding-box $\mathcal{B}$. A set of scores S contains the likelihoods of the presence of an instance of each known class within the limits of $\mathcal{B}$. Notice that S does not need to be normalized.

In a nutshell, *SemBA* maps a scene by updating its current semantic beliefs $\boldsymbol{\beta}$ with likelihoods S from new observations. These observations are sequentially extracted from each fixation f_t along the scanpath. Each belief $\boldsymbol{\beta}^{\mathbf{x}}$ is associated with a cell $\mathbf{x} = (x, y)$, where $x \in \{1, \ldots, X\}$ and $y \in \{1, \ldots, Y\}$. *SemBA*'s active perception mechanism for visual search aims at shifting the gaze to a cell $\mathbf{x}$ that maximizes a posterior categorical probability $P\left(C^{\mathbf{x}} = k\right)$ for a target class k. Dirichlet distributions serve as conjugate prior distributions of both categorical and multinomial distributions. Taking advantage of this property, posteriors are modeled through sets of Dirichlet parameters $\boldsymbol{\beta}^{\mathbf{x}} \in \mathbb{R}^K$, already introduced as semantic beliefs. To define an initial state of maximum entropy, all beliefs are initialized as $\beta_k^{\mathbf{x}} = 1, \forall k \in \{1, \ldots, K\}$, conforming to flat Dirichlet distributions that represent non-informative priors.

Each time a detection's bounding-box $\mathcal{B}$ overlaps the region that corresponds to a cell $\mathbf{x}$, the beliefs of that cell are updated:

$$\beta_k^{\mathbf{x}} \longleftarrow \frac{\beta_k^{\mathbf{x}} \left(1 + \frac{s_k}{\sum_{j=1}^{K} \beta_j^{\mathbf{x}} s_j}\right)}{1 + \frac{\min_i s_i}{\sum_{j=1}^{K} \beta_j^{\mathbf{x}} s_j}}, \forall k \in \mathcal{C}, \tag{1}$$

with $s_k \in S$ representing that detection's categorical likelihood for each class $k \in \mathcal{C}$. This update rule, borrowed from Kaplan's work [20] on classifier fusion, applies subjective logic to incorporate and manage uncertainty in the scores S.

Baseline (Full Image) | ***Laplacian Foveation*** | ***FOVEA* (Magnification)** | ***Multi-Scale Fovea***

Fig. 3. Comparison of the field-of-view topologies for each different foveal system used in this work: Full resolution image (baseline), *Laplacian Foveation* [14], *FOVEA* (Magnification) [17], and our *Multi-Scale Fovea* (in order, from left to right). Here, *Multi-Scale Fovea*'s parameters are set such that the maximum acuity region, corresponding to the fovea [6], comprises a region equivalent to around 2.0° of visual angle (64×64 pixel patch in a 1050×1680 image). Note that *SemBA* processes each *Multi-Scale Fovea* layer separately. Here, to showcase the topology, we overlap all layers while preserving their resolution.

Finally, to determine the next-best viewpoint $\mathbf{x}^*$ at time t, *SemBA* greedily selects the cell $\mathbf{x}$ with maximum expectancy

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmax}}\, \mathbb{E}\left[C^{\mathbf{x}} = k \mid \boldsymbol{\beta}^{\mathbf{x}}\right], \tag{2}$$

where $\mathbb{E}\left[C^{\mathbf{x}} = k \mid \boldsymbol{\beta}^{\mathbf{x}}\right] = \beta_k^{\mathbf{x}} / \sum_{j=1}^{K} \beta_j^{\mathbf{x}}$. The gaze is then shifted to the next-best viewpoint $f_{t+1} \equiv \mathbf{x}^*$ (in pixels), according to the information gathered from $f_{0:t}$ and fused in $\boldsymbol{\beta}^{\mathbf{x}}$. *SemBA* iteratively repeats this process, of collecting and fusing new detections from each new f_{t+1}, until some terminal criterion is met, while applying inhibition of return (IOR [27]).

C. Multi-Scale Fovea Module

Drawing inspiration from the seminal work of Bandera and Scott [12], we now describe the implementation of our foveal mechanism for efficient bio-inspired semantic data extraction.

Let us start by considering a total of N layers $L_1, \ldots, L_N$. Each layer L_n consists of a square-shaped crop, centered on a focal point $f \in \mathbb{N}^2$, with sides of length $l_n \in \mathbb{N}$ (in pixels). Hence, a level L_n can be represented by the coordinates of its top-left $(-)$ and bottom-right $(+)$ coordinates in the main image frame, i.e. $L_n = [c_n^-, c_n^+]$ such that $c_n^\pm \in \mathbb{N}^2$, similar to a bounding-box. Therefore, the corner coordinates of each layer L_n depend on $f = (x_c, y_c)$ and l_n as: $c_n^\pm = (x_c \pm l_n/2,\, y_c \pm l_n/2)$. The pseudo-foveal region is assumed to be contained within a base layer L_1 with an assigned side length l_1. To allow the fovea to move close to the scene's boundaries, we apply zero padding to the original image. The side lengths for subsequent layers are then defined as $l_{n+1} = 2l_n$, such that $l_n = 2^{n-1} l_1, \forall n \in \{1, \ldots, N\}$. After being cropped, all layers are then downsampled to the size of the layer L_1 via bilinear interpolation. However, L_1 preserves its original dimensions, retaining maximum acuity, just as the human fovea. This process, known as exponential pixel density roll-off [12], more intensely degrades the visibility of objects rendered in the outermost layers, akin to peripheral distortion.

All layers L_n are then fed to a deep object detection architecture [23]–[25] to extract their semantic content. Detections gathered from each layer must first be remapped to the original image $Y \times X$ grid and only then appropriately fused (1) using Kaplan's rule. Let us consider an arbitrary detection from a layer L_n, with its corresponding bounding box $\mathcal{B}' = \left[p'_-, p'_+\right]$, represented by its coordinates in the layer frame: $p'_\pm = \left(x'_\pm\, y'_\pm\right)$. Finally, we transform the coordinates of $\mathcal{B}'$ to the main image frame, and obtain $\mathcal{B} = [p_-, p_+]$ as

$$p_\pm = \left(x_c - \frac{l_n}{2} + x'_\pm 2^{n-1}, y_c - \frac{l_n}{2} + y'_\pm 2^{n-1}\right). \tag{3}$$

Kaplan's rule (1) is applied to each detection obtained from every single layer, such that multiple observations, potentially corresponding to the same object, are fused sequentially within a single fixation f_t. Because this fusion rule promotes low-variance estimates between consecutive updates [20], aggregating a larger number of high-confidence observations at a given location results in increased posterior confidence for the associated category. Objects nearer to the fovea's center are processed across a greater number of scales, yielding more corresponding observations. As expected, this leads to more confident estimates for objects displayed closer to the fovea and more uncertain estimates for objects located further into the periphery, mirroring the human visual-cognitive system.

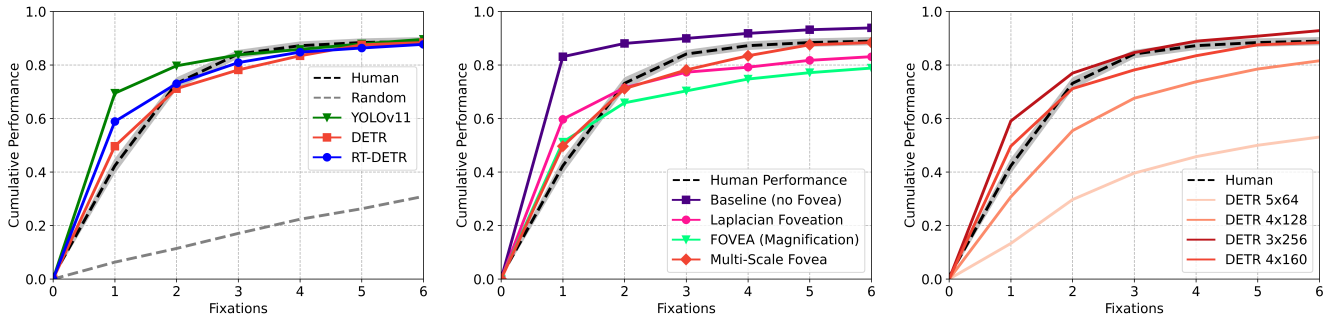
Since $N \times l_1 \times l_1 << \textit{height} \times \textit{width}$, we can assert that this foveal mechanism effectively reduces the total amount of pixels to be processed by any given detector, provided that the base layer size is small enough, i.e. $l_1 < \min\{\textit{height}, \textit{width}\}$.

IV. Experiments and Results

A. Experimental Setup

To appropriately assess the performance of the proposed *Multi-Scale Fovea* (incorporated into *SemBA* [10]) in target-present visual search, we use off-the-shelf sequence matching metrics [32]–[34]. To comparatively evaluate the impact of the proposed foveal system, we will consider two other types of foveated fields-of-view: *Laplacian Foveation* [14] and *FOVEA* [17]. Moreover, given that *SemBA* heavily relies on efficient and accurate object recognition [9], we conduct ablation studies using three distinct state-of-the-art object detection models: DETR [24], YOLOv11 [23], and RT-DETR [25].

COCO-Search18 [27] is a subset of the well-known COCO 2017 [26] dataset, comprising a total of 6202 images, 3101 target-present (TP) and 3101 target-absent (TA) scenes, each associated with one of the 18 possible target object categories. For each instance, whether TP or TA, there were collected the



(a) Deep Object Detection Models (b) Artificial Foveal Systems (c) *Multi-Scale Fovea Settings*

Fig. 4. Cumulative performances of humans (average), a random selection algorithm, and *SemBA* under different (a) object detectors, (b) foveal systems, and (c) *Multi-Scale Fovea* configurations, on target-present visual search. Regarding experiment (a) we apply our *Multi-Scale Fovea* on a 4×160 configuration. We use the same configuration for experiment (b) when assessing the performance of the *Multi-Scale Fovea*. For the experiments (b) and (c) we use *SemBA*$\times$DETR.

natural scanpaths of 10 human subjects, using modern eye-tracking technology. Besides the TP/TA setting distinction, the dataset is split into 3 partitions: train, validation and test sets. To promote a fair comparison with other scanpath prediction models, we evaluate each method on COCO-Search18's test partition alone. COCO-Search18's test set comprises COCO 2017 train and validation samples. Because most deep object detection models [23]–[25] are pre-trained on the full COCO 2017 dataset, we are forced to retrain the models, without the samples that are part of COCO-Search18, to prevent information leakage. Therefore, we retrain YOLOv11, DETR, and RT-DETR for 300 epochs using default hyperparameters.

The models are trained to localize and classify 80 distinct object categories. To ensure that objects located on the blurry peripheral region of the field-of-view are also detected, we set the confidence threshold to $1\,\%$ for all models. By fusing such ambiguous semantic scores we intentionally add uncertainty to the attention mapping process. DETR's architecture [24] uses ResNet-50 as backbone, adding dilation and removing a stride from the first convolution of its last stage. We also consider large YOLOv11 and RT-DETR (HGNetv2) implementations [23], containing 25.3 and 32 million parameters, respectively.

As illustrated in Fig. 2, *SemBA* splits an arbitrary size image into a 20×32 grid. A fixation point is defined as the center-most pixel of a selected cell. After each fixation, IOR is applied in a 3×3 block [27], centered on the cell that was being fixated. This corresponds to a visual angle of 5.0° in diameter, which is similar to the area covered by the parafovea [7]. When dividing a 1050×1680 COCO-Search18 image into a 20×32 grid, a 3×3 block of cells covers roughly a 160×160 pixel patch. Hence, when comparing with other scanpath prediction models (Tab. I), *SemBA*'s fovea module builds a 4 level pyramid with base length $l_1 = 160$. To assess fovea's size impact on search accuracy and human scanpath similarity (Tab. II) we experiment with *SemBA* and *Multi-Scale Fovea* assigning larger ($l_1 = 256$) and smaller ($l_1 = 64$ and $l_1 = 128$) dimensions to the layers. The number of levels is chosen so that each configuration encapsulates the exact same region: $N = 3$ for $l_1 = 256$, $N = 4$ for $l_1 = 128$, and $N = 5$ for $l_1 = 64$, covering a total area of 1024×1024 pixels. Such settings ensure that the majority of semantic information is available (i.e. displayed in the field-of-view) when fixating on the center of any 1050×1680 image. Both *Laplacian Foveation* [14] and *FOVEA* [17] topologies follow the configurations illustrated in Fig. 3. Due to its high sensitivity to the choice of foveal proportions, we set the radius of the *Laplacian Foveation* to be the diagonal length of a 3×3 block of cells. This choice makes the effective foveal area comparable to that of a 3×256 *Multi-Scale Fovea* configuration. Regarding *FOVEA*, we apply a low-variance Gaussian distribution, centered around the focal point, to generate the warped field-of-view, which compresses the visual appearance of objects located in the peripheral region.

In each experiment, we generate scanpaths for 586 samples from COCO-Search18's test set, in their maximum resolution (1050×1680). By benchmark convention [27], search is always initiated (f_0) at the center of each test image. Similar to IVSN [31], we apply an oracle that stops the search process once the gaze is set upon the targeted object's ground-truth bounding-box. By opting for an oracle, we avoid the need to tackle the challenging problem of defining a hard confidence threshold τ that terminates the search process, i.e. $\mathbb{E}\left[C^{\mathbf{x}} = k \mid \boldsymbol{\beta}^{\mathbf{x}}\right] \geq \tau$.

B. Evaluation Metrics

We now describe the off-the-shelf benchmark metrics [33] used for assessing the similarity between ground-truth human scanpaths and model-generated fixation sequences. Sequence Score (**SS**) represents scanpaths as ordered strings of fixation cluster IDs and quantifies their similarity through the Needleman–Wunsch global alignment algorithm [29]. This algorithm was originally designed for comparing between two protein amino-acid sequences. In contrast, Fixation Edit Distance (**FED**) encodes scanpaths using the same cluster-based string representation but evaluates dissimilarity via Levenshtein's edit distance [30]. Semantic Sequence Score (**SemSS**) departs from SS by replacing spatial cluster identifiers with categorical semantic labels that denote the objects fixated at each step, while retaining Needleman–Wunsch's algorithm. Analogously,

TABLE I
SCANPATH PREDICTION MODEL EVALUATION IN VISUAL SEARCH

	SemSS ↑	SemFED ↓	SS ↑	FED ↓
Human Consistency	0.470	2.144	0.463	2.353
IVSN [31]	0.380	3.034	0.346	3.360
Gazeformer [32]	0.490	1.928	**0.504**	2.072
HAT [33]	0.540	1.522	0.468	2.063
CLIPgaze [34]	**0.545**	**1.489**	0.476	**2.014**
SemBA × YOLOv11*	0.448	2.161	0.414	2.572
SemBA × DETR*	0.488	2.102	0.426	2.574
SemBA × RT-DETR*	0.473	2.167	0.421	2.635

*Incorporating our *Multi-Scale Fovea* in a 4×160 pixel configuration.

TABLE II
ABLATION STUDY AND COMPUTATIONAL COST EVALUATION

	SemSS* ↑	SemFED* ↓	Pixels ↓
Human Consistency	0.470	2.144	-
Baseline (no Fovea)	0.430	**2.025**	100 %
Laplacian Foveation [14]	0.416	2.488	100 %
FOVEA (Magnification) [17]	0.384	2.769	100 %
Multi-Scale Fovea 5×64	0.351	3.771	0.01 %
Multi-Scale Fovea 4×128	0.460	2.409	3.72 %
Multi-Scale Fovea 3×256	0.464	2.099	11.1 %
Multi-Scale Fovea 4×160	**0.488**	2.102	5.80 %

*Original image resolution of 1050×1680, applying *SemBA*×DETR.

Semantic Fixation Edit Distance (**SemFED**) operates on semantic label sequences but adopts Levenshtein's edit distance.

To quantify human consistency, as reported in Tab. I and Tab. II, we compute the four distance metrics (SS, FED, SemSS, SemFED) between all pairs of scanpaths (from 10 human subjects) for each COCO-Search18 test image. For each considered trial, we ignore whether the subject's response was correct or incorrect. Table entries correspond to the mean value of each metric, averaged over all possible scanpath pairs. Moreover, to assess task performance, we also plot the cumulative performance (Fig. 4) attained by each model or foveal setting. It consists of the ratio between the number of samples where the target was already found and the total number of test samples, after a certain number of fixations.

To assess model performance [28], we compare the scanpath generated by the model with the scanpaths generated by human subjects for the same test sample, averaging the respective metric scores. To promote a fair comparison [9], we truncate the lengths of all scanpaths to 6 fixations (excluding f_0). For each metric, we highlight the best model's value in bold, while underlining other models' values that beat human consistency.

C. Results and Discussion

In Fig. 4 we show the cumulative performance of *SemBA* in target-present visual under different settings. For comparison, we present the results of a random gaze selection method as well as the average human performance (across 10 subjects) together with the respective standard error of the mean bands.

Setting our *Multi-Scale Fovea* parameters to cover a $5.0°$ angle (4×160 configuration), we assess our attention pipeline's performance when incorporating different (a) object detection and (b) foveal mechanism modules. From the curve presented in Fig. 4 (a) we observe that, under the mentioned fovea setup, *SemBA* achieves the human average cumulative performance after 6 fixations (around $90\,\%$), whether paired with YOLOv11, DETR, or RT-DETR. However, regarding the performance after just 1 fixation, YOLOv11, RT-DETR, and DETR (about $70\,\%$, $60\,\%$, and $50\,\%$, respectively) tend to overshoot human results (near $40\,\%$). Nevertheless, DETR's performance curve is clearly the most similar to the human curve. Therefore, we select DETR as our go-to model when assessing the artificial fovea module configurations in Fig. 4.

Regarding scanpath similarity, Tab. I reveals that *SemBA* is able to level human consistency in terms of sequence semantic similarity (SemSS and SemFED), even slightly surpassing it. However, *SemBA* is still below human consistency in terms of sequence location distance (SS and FED). On the one hand, given that *SemBA* relies solely on semantic information, it is critical to conclude that produced sequences of fixated objects (both target and distractors) accurately mimic human patterns. On the other hand, the fact that *SemBA* does not learn from actual human scanpaths possibly hinders its fixation location accuracy when compared to other benchmark models [32]–[34]. However, with the new *Multi-Scale Fovea*, *SemBA* still outperforms [9] the best performing scanpath prediction model that does not learn from human scanpaths, i.e. IVSN [31]. When analyzing semantic distance metrics, *SemBA* is the model that best approximates human consistency, despite underperforming compared to state-of-the-art models [32]–[34], which substantially overshoot inter-human performance.

With respect to the evaluation of the artificial foveation module's impact on *SemBA* performance, using DETR, Fig. 4 (b) shows that, for the selected parameters, the *Multi-Scale Fovea* obtains the best task performance after 6 fixations. Furthermore, its curve better approximates the human curve when compared to the other foveal topology curves. Regarding semantic distance metrics, Tab. II reveals that only the newly proposed fovea is able to surpass human consistency, beating both *Laplacian Foveation* and *FOVEA*. As highlighted, the key advantage of the *Multi-Scale Fovea* is that it heavily reduces the amount of pixels to be processed by the object detector. Note that *FOVEA* (and possibly *Laplacian Foveation*) could eventually yield more accurate results if the selected object detector was retrained after applying the respective foveal transforms to the training dataset. However, because our *Multi-Scale Fovea* preserves critical geometric properties of regular images, we avoid the need to retrain object detectors, which is a very time-consuming and computationally costly process.

Finally, we conduct an ablation study to determine the influence of the *Multi-Scale Fovea* parameters. As expected, Fig. 4 (c) conveys that decreasing the fovea dimensions generates a lower cumulative performance due to the increased peripheral distortion, and vice versa. As previously mentioned the 4×160 configuration, which is more congruent with the actual fovea

anatomical proportions, better approximates the human curve while the 3×256 setting essentially serves as its upper-bound. Moreover, unlike in the other configurations, only in the 4×160 setting does *SemBA* excel human consistency across both semantic sequence similarity metrics, as shown in Tab. II.

To assess whether our novel *Multi-Scale Fovea* effectively promotes computational gains in object detection, we run inference (NVIDIA GeForce RTX 4060) on a full 1050×1680 image and a 4×160 setting, using YOLOv11 and DETR. In terms of time cost (in seconds), *SemBA* $\times$ DETR drops from $10.37\,\mathrm{s}$ to $0.59\,\mathrm{s}$ per iteration, achieving a notable 17.6x speed-up. However, *SemBA* $\times$ YOLOv11 only drops from $0.144\,\mathrm{s}$ to $0.115\,\mathrm{s}$, showing a minor speed-up. This results show that the proposed *Multi-Scale Fovea* considerably diminishes detection costs for heavier models (e.g. DETR), but promotes minimal gains when applying modern architecture (e.g. YOLOv11) that are by now prepared to swiftly handle larger visual inputs. Nevertheless, modern architectures, which generally achieve higher levels of accuracy, appear to inadequately capture the uncertainty that is inherently linked to the attentional behavior.

V. Conclusions

Deep object detection models excel at extracting top-down semantic cues [18], which constitute one of the main sources of preattentive information. Because the size of the visual input has been traditionally considered a bottleneck in object detection [13], we propose a *Multi-Scale Fovea* mechanism that reduces the total amount of pixels to be processed from each fixated point. Inspired by the anatomy of the human visual system, our novel method builds a multi-resolution pyramid [12], around a focal point, which gradually degrades the quality of information in more peripheral levels. We show that the *Multi-Scale Fovea* is able to improve the performance [9] of our semantic-based Bayesian attention framework (*SemBA*) in target-present visual search. The proposed hard-attention mechanism leads *SemBA* to more closely mimic human gaze patterns, notably in terms of the sequences of fixated object categories. Therefore, we show that the relationship between overt and covert attention can be effectively modeled within an efficient framework that learns solely from cues extracted directly from visual *stimuli*. As future work, we aim to extend the proposed framework and foveal system to other visual tasks, namely free-viewing and visual question answering.

References

[1] J. Wolfe. "Guided Search 6.0: An updated model of visual search," in Psychonomic bulletin & review, vol. 28, no. 4, pp. 1060–1092, 2021.

[2] 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.

[3] L. Itti, C. Koch, et al., "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, 1998.

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

[5] 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.

[6] W. S. Tuten and W. M. Harmening, "Foveal vision," Current Biology, vol. 31, no. 11, pp. R701–R703, 2021, doi: 10.1016/j.cub.2021.03.097.

[7] E. Stewart, M. Valsecchi, A. Schutz. "A review of interactions between peripheral and foveal vision," in Journal of vision, vol. 20, no. 12, 2020.

[8] C. F. Staugaard, A. Petersen, and S. Vangkilde, "Eccentricity effects in vision and attention," Neuropsychologia, vol. 92, pp. 69–78, 2016.

[9] J. Luzio, et al., "Human Scanpath Prediction in Target-Present Visual Search with Semantic-Foveal Bayesian Attention," in 2025 IEEE International Conference on Development and Learning (ICDL), 2025.

[10] J. Luzio, A. Bernardino, and P. Moreno, 'SemBA-FAST: Semantic-based Bayesian attention applied to foveal active visual search tasks', Neurocomputing, vol. 673, p. 132860, 2026.

[11] J. W. Bisley, "The neural basis of visual attention," The Journal of physiology, vol. 589, no. 1, pp. 49–57, 2011.

[12] C. Bandera, P. Scott, "Foveal machine vision systems," in Conference Proceedings., IEEE International Conference on Systems, Man and Cybernetics, pp. 596-599 vol.2, 1989.

[13] F. Arrebola, P. Camacho, F. Sandoval, "Generalization of shifted fovea multiresolution geometries applied to object detection," in Image Analysis and Processing, pp. 477–484, 1997.

[14] A. F. Almeida et al., "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.

[15] P. Ozimek, et al. "A space-variant visual pathway model for data efficient deep learning," Frontiers in cellular neuroscience, vol. 13, pp. 36, 2019.

[16] 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.

[17] C. Thavamani, et al., "Fovea: Foveated image magnification for autonomous navigation," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 15539–15548, 2021.

[18] P. Tsirtsakis et al., "Deep learning for object recognition: A comprehensive review of models and algorithms," International Journal of Cognitive Computing in Engineering, vol. 6, pp. 298–312, 2025.

[19] T. Shehzadi, K. A. Hashmi, et al., "Object detection with transformers: A review," Sensors, vol. 25, no. 19, p. 6025, 2025.

[20] L. Kaplan, et al., "Fusion of classifiers: A subjective logic perspective," 2012 IEEE Aerospace Conference, Big Sky, MT, USA, pp. 1-13, 2012.

[21] Joseph Redmon et al., "You only look once: Unified, real-time object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016.

[22] J. Terven et al., "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas," Machine Learning and Knowledge Extraction, vol. 5:4, pp. 1680–1716, 2023.

[23] G. Jocher and J. Qiu, Ultralytics YOLO11. 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[24] 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, pp. 213–229, 2020.

[25] Y. Zhao et al., "DETRs Beat YOLOs on Real-time Object Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16965–16974, 2024.

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

[27] 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.

[28] 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, pp. 770–787, 2018.

[29] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," Journal of molecular biology, vol. 48, no. 3, pp. 443–453, 1970.

[30] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Proceedings of the Soviet physics doklady, 1966.

[31] 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.

[32] S. Mondal, Z. Yang, S. Ahn, et al., "Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention," CVPR, pp. 1441–1450, Jun. 2023.

[33] Z. Yang, et al., "Unifying Top-down and Bottom-up Scanpath Prediction Using Transformers," CVPR, Jun. 2024.

[34] Y. Lai et al., "CLIPGaze: Zero-Shot Goal-Directed Scanpath Prediction Using CLIP," in ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5, 2025.