具有语义-中央凹贝叶斯注意的目标存在视觉搜索中的人类注视路径预测

João Luzio, Alexandre Bernardino, and Plinio Moreno

***Abstract*— 在目标导向的视觉任务中，人类的知觉受到自上而下的线索和自下而上的线索的引导。同时，中央视觉在高效引导注意力中起到关键作用。现代对仿生计算注意力模型的研究利用了深度学习的进步，通过利用人类注视路径数据实现了新的最先进性能。在这项工作中，我们评估了 SemBA -FAST 的性能，即用于中央主动视觉搜索任务的基于语义的贝叶斯注意力这一自上而下框架，旨在预测目标存在下的人类视觉注意力。SemBA -FAST 将深度物体检测与概率语义融合机制集成起来，动态生成注意力图，利用预训练的检测器和人工中央凹来更新自上而下的知识并顺序地提高注视预测。我们在 COCO-Search18 基准数据集上对 SemBA -FAST 进行了评估，将其性能与其他注视路径预测模型进行比较。我们的方法达成了与人类真实注视路径紧密匹配的注视序列。值得注意的是，它超越了基准线和其他自上而下的方法，在某些情况下与注视路径知情模型竞争。这些发现为语义-中央凹概率框架在人类模拟注意力建模能力上提供了宝贵的见解，并对实时认知计算和机器人技术有重要意义。**

I. 介绍

视觉注意力受人类眼动系统和认知系统的双重影响。一方面，人类的视觉传感器（眼睛）作为一种硬注意力的感官机制 [?] ，动态地限制着视觉场中内容的可见性。另一方面，由于可用的视觉认知系统的脑资源有限，人类的主动感知机制 [?] 被迫有效利用可用的信息。

由人类视觉感官系统施加的主要解剖学限制称为中央凹视力 [?] ，它有效地减少了在每次注视停留期间需要处理的总信息量。这是通过明显保留视野的中心清晰区域，即称为中央凹的部分，同时逐渐增加其周围区域的模糊程度，通常称为外围视野。与此同时，一个隐蔽的注意机制 [?] 处理感知到的信息，以确定下一个注视点，旨在关注任务相关性最显著的区域。

总体而言，人类视觉注意依赖于两种类型的信息信号：自底向上和自顶向下。自底向上的显著特征可以直接从视觉刺激中提取 [?] ，而自顶向下的信息显著性是目标导向的，取决于实际任务的性质。自底向上的线索 [?] 通常由颜色和亮度强度对比或边缘和图案的几何方向触发。自顶向下的指导 [?] 来自于先验知识、任务需求、期望以及目标等因素。

在人类常执行的众多视觉认知任务中，有两个任务在注意力建模研究领域尤为受到关注：自由观视和视觉搜索。前者指的是无特定目标地自由探索一个场景。因此，自由观视主要由自下而上的低级特征驱动。而后者是指在可能包含多个干扰物的视觉场域中搜索给定目标类别的实例。由于其目标导向的特性，视觉搜索高度依赖如场景上下文和语义等自上而下的特征。

João Luzio, Alexandre Bernardino, and Plinio Moreno are with the Institute for Systems and Robotics, Instituto Superior Técnico, University of Lisbon, Portugal. Email: joaoluzio14@tecnico.ulisboa.pt

This work was supported by *Fundação para a Ciência e Tecnologia*.

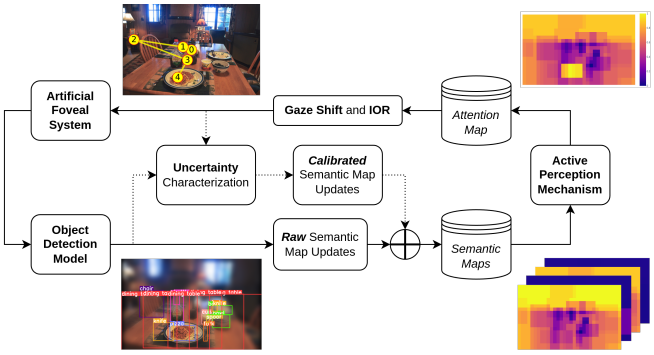

Fig. 1. SemBA-FAST [?] 的方法流程，受 [?] 启发。中央注视图像被送入一个深度目标检测模型，该模型生成多个边界框及其相应的类别分数。该信息用于更新固定在世界坐标系上的语义地图。可以通过两种方法进行更新：使用目标检测器的原始分数（实线）或根据中央注视图像特征的影响校准的分数（虚线），因为外围的模糊增加会增加分类分数的不确定性。我们的主动感知机制利用在语义地图中收集和存储的自上而下的信息来构建一个单一的注意力地图。我们从生成的注意力地图中推断出下一个最佳视图，并应用返回抑制（IOR [?], [?], [?] ）机制以防止重复搜索先前的位置。

视觉搜索实验通常在两种不同的设置下进行 [?] ：目标存在（TP），其中感知到的视觉刺激包含至少一个目标类别的实例；以及目标不存在（TA），即在视野范围内没有所搜索类别的对象存在时的情况。

理解人们通过预测其扫视路径（即凝视顺序）寻找目标的方法，对于人机交互和开发能够预测用户需求和意图的系统具有重要意义。人类视线预测在诸多应用中具有极高的价值，例如早期诊断认知障碍 [?] 和心理健康状况 [?] ，以及类人机器人 [?] ，甚至中央凹渲染 [?] 。

由于深度学习模型的最新突破以及人类注视数据集和基准的建立，关于人类注意力预测的研究已经从高层特征提取方法（例如 [?] ）转向使用实际人类扫视路径数据训练的模型（例如 [?], [?], [?] ）。尽管像 Gazeformer [?] 、HAT [?] 和 CLIPgaze [?] 这样的模型极大地提升了最新技术的性能，它们最终缺乏模仿视觉-认知系统行为的方法（例如 [?] ）的生物学合理性和可解释性。例如，人类并不通过同伴的经验和数据来学习如何探索其周围环境或完成复杂的视觉任务。实际上，尽管需要诸如语义和上下文的自上而下的知识 [?] ，幼儿在学习如何探索复杂拥挤场景以迅速找到目标物体时，并未接收到任何类型的扫视路径数据。事实上，人类凭直觉学习这些技能 [?] ，因为我们内置的视觉-认知系统有效地引导我们的注意力，使我们能够更好地感知和与周围环境互动。

在 IVSN [?] 之后，一种主要从提取的自上而下特征构建注意力图的零样本模型，没有发布过很多纯粹由刺激驱动的模型。受人眼的解剖特性和大脑注意力图组装过程的启发 [?] ，我们提出了一种基于语义的贝叶斯注意方

法论 [?] ，我们恰当地命名为 SemBA。我们遵循神经科学文献中的常见趋势 [?], [?] ，将中心视角和语义数据融合结合到一个单一的架构中。

在这项工作中，我们专注于 TP 视觉搜索。我们旨在评估我们提出的用于中央凹主动搜索任务的自上而下概率注意框架，即 SemBA -FAST [?] ，是否能准确预测人类生成的扫描路径。我们的模型和方法如图 1 所示，灵感来源于 Dias 等人的工作 [?] ，该工作关于场景探索的语义-中央凹主动感知（即自由观看）。这项具体工作的主要贡献如下：

- 我们在一个知名的视觉搜索基准，即 COCO-Search18 [?] 数据集上，使用现成的评估标准 [?] ，对比了 SemBA-FAST 的 [?] 实验结果与其他最先进模型 [?], [?], [?], [?], [?], [?], [?] 的结果。
- 作为一种纯语义模型，SemBA -FAST 已被证明在 TP 视觉搜索任务中达到与人类结果 [?] 相当的综合性能。
- 我们通过实验展示了 SemBA -FAST 能够自信地超越基线模型和其他自上而下的方法 [?] 。此外，研究的模型在某些注视预测指标方面也能超越基于人类注视数据的信息模型 [?], [?], [?], [?], [?], [?] 。

II. 背景和相关工作

A. 深度目标检测

目标识别是一个著名的计算机视觉问题，涉及在图像和视频帧中定位（通过边界框确定位置）和分类（确定相应类别）可见的物体。最初这一领域被传统算法所主导，但存在诸多限制 [?] ，随着深度学习模型的发展，该领域经历了显著的变革。

在目标检测研究领域 [?] ，模型分为两阶段检测器和单阶段检测器。两阶段检测器（例如 R-CNN [?] ）首先提出感兴趣区域，然后才对这些区域进行分类。相比之下，单阶段检测器（例如 YOLO [?] 及其后续版本，即 YOLOv2 到 YOLOv11 [?] ）省略了感兴趣区域的搜索阶段，直接预测物体的位置和类别，这往往会提高训练和推理的速度。最近，基于 transformer 的架构（例如 DETR [?] ）在目标识别领域中颠覆了最新技术，将任务视为直接的集成预测问题。

一般来说，单阶段检测器在速度方面已被证明优于双阶段检测器，并且其准确性也在不断提高。YOLO 系列 [?] 不断发展，成为实时目标识别中最流行和最有效的检测器类别之一。尽管有这些进展，像在可变光照条件下的检测和物体遮挡等挑战仍然存在。

B. 注意建模

理解人类如何控制和引导他们的凝视运动 [?] 是理解我们注意力引导的视觉行为模式的基础。虽然视觉注意力预测在数十年来一直是神经学家和心理学家感兴趣的领域 [?], [?] ，但这一主题最近才引起计算机视觉和深度学习研究社区的关注。L. Itti 和 C. Koch 在低级特征提取用于视觉显著性映射的基础性工作 [?] 促进了许多后续的底层显著性分析的发展，以用于人类注意力建模 [?] 。

除了忽略自上而下与任务相关的特征 [?] 外，纯粹基于自下而上的显著性注意研究仅专注于区分整个视野中不同区域的显著性水平。虽然显著图可能有效地模拟注视点的空间分布，但它们几乎不能提供关于预测焦点的顺序的见解。出于这些原因，人类视线路径的预测对传统注意力建模提出了真正的挑战，因为我们不仅对预测注视位置感兴趣，还对注视的时间顺序感兴趣。

C. 人类视线路径预测

我们已经断言，目标导向的注视预测依赖于提取和处理自上而下特征的能力。Zhang 等人利用他们的自上而下零样本模型（称为不变视觉搜索网络，IVSN）的提议，极大地丰富了受生物启发的扫描路径预测研究。零样本搜索包括寻找未出现在训练数据集中的类别实例。

在 IVSN 之后，随着首个大规模视觉搜索基准数据集 COCO-Search18 [?] 的建立，众多基于深度学习的人类视线轨迹预测模型 [?] 得到了推动。利用眼动追踪技术，COCO-Search18 提供了许多由 10 名被试生成的注视序列，包括目标出现和目标缺失两种设定。在训练过程中结合视线轨迹，诸如 IRL [?] （逆强化学习）和 FFMs [?] （黄斑特征图）等模型已经能够超越 IVSN [?] 和其他最先进的基准模型 [?] 的表现。陈等人 [?] 同样成功地将基于逆强化学习的视线轨迹预测扩展到另一项具有挑战性的任务，即视觉问答（VQA）。更近期，Gazeformer [?] 和 HAT [?] 能够利用软注意机制，这是基于 Transformer 架构（ViT）的特点，在 TP 和 TA 视觉搜索任务中超越所有竞争对手。两个模型均从人类视觉系统中汲取灵感，通过模拟简化的黄斑视网膜实现动态视觉记忆。最后，CLIPgaze [?] 利用大型视觉语言模型（VLM），即 CLIP，为图像和其相应目标提示提取并提供预匹配特征，以便在 TP、TA 以及零样本设定下搜索对象。

III. 方法论

在本节中，我们描述了我们的人类注意力预测方法的完整流程：SemBA，即基于语义的贝叶斯注意力。在一个利用预训练目标检测器的概率框架内，我们考虑了一种用于自顶向下的语义信息融合的机制 [?] 。我们在图 1 中展示了我们方法的示意图。

A. 空间约束

与其他最新的人类扫描路径预测方法类似 [?], [?], [?] ，我们将二维图像视为固定的视野，在这种视野中，场景的空间配置和边界不能动态改变。此外，我们假设视觉字段具有静态配置 [?] ，其中包含的物体的位置随时间不变。考虑到这一假设，我们在像素级定义每个注视的坐标，以减轻与网格离散化相关的精度损失 [?] 。考虑一个尺寸为 $height \times width$ 的图像和一个给定的初始注视点 f_0 ，通常设在视野中心。我们的模型生成一系列类似人类的注视点 $f_1, f_2, \ldots, f_n$ ，其中每个注视 $f_t, \forall t \in \{1, \ldots, n\}$ 对应于图像内的特定像素位置。每个注视序列的长度 n 因场景不同而变化，因为它取决于设定的终止准则 [?] 。对于 SemBA 的输出语义和注意力图（见图 1 ），我们选择一种典型的二维笛卡尔表示，在空间上对周围环境编码，以 $Y \times X$ 上下文网格的形式。通过动作空间离散化，我们旨在降低计算成本 [?] ，并通过处理更广泛的兴趣区域而非单个像素来接近人类认知系统的敏感度级别 [?] 。

B. 概率框架

如图 1 所示，每个检测由一个边界框和一个维度为 K 的得分向量 S 组成，该得分向量通常是规范化的。向量

S 包含一组置信度，每个置信度与其边界框内一个特定已知类别实例的存在相关联：

$$S = (s_1, s_2, \ldots, s_K), 0 \leq s_k \leq 1, \forall k \in \mathcal{C} \tag{1}$$

其中 $\sum_{i=1}^{K} s_i = 1$ 和 $\mathcal{C} \subseteq \{1, \ldots, K\}$ 表示检测器已知的类别集合。位于 $\boldsymbol{x} = (x, y)$ 的物体的类别被建模为参数为 $\boldsymbol{\beta}^{\boldsymbol{x} \in \mathbb{R}^K}$ 的 Dirichlet 复合多项式分布（信念）。

$$P\left(C^{\boldsymbol{x}} = k \mid \boldsymbol{\beta}^{\boldsymbol{x}}\right) = \frac{n!}{\left(\sum_{i=1}^{K} \beta_i^{\boldsymbol{x}}\right)^n} \prod_{i=1}^{K} \frac{(\beta_i^{\boldsymbol{x}})^{n_i}}{n_i!} \tag{2}$$

其中 $n = \sum_{i=1}^{K} n_i, n_k = 1$ 和 $n_i = 0, \forall i \in \mathcal{C} \setminus k$ [?] 。我们为每个类别 $k \in \mathcal{C}$ 构建语义地图，如图 1 所示，通过对所有网格位置 $\boldsymbol{x}$ 的 (2) 扩展。

Dirichlet 信念集的参数初始化为 $\beta^{\boldsymbol{x}_k = 1, \forall k \in \{1, \ldots, K\}}$ ，以定义与非信息先验相对应的平坦 Dirichlet 分布。

本质上，我们的方法论的目标是使用来自新观察的得分 (1) 来更新 Dirichlet 复合多项式分布的 $\boldsymbol{\beta}^{\boldsymbol{x}}$ 参数，这些得分是从扫描路径中的每一个凝视点依次提取出来的。

C. 语义数据融合

更新语义地图最直接的方法是使用分类器输出得分向量，如图 1 中实线路径所示。为了涵盖多次凝视中获得的不同得分，我们应用了一种从主观逻辑视角发展而来的分类器融合规则 [?] ，在规则

$$\beta_k^{\boldsymbol{x}} \longleftarrow \frac{\beta_k^{\boldsymbol{x}} \left(1 + \frac{\lambda_k}{\sum_{j=1}^{K} \beta_j^{\boldsymbol{x}} \lambda_j}\right)}{1 + \frac{\min_i \lambda_i}{\sum_{j=1}^{K} \beta_j^{\boldsymbol{x}} \lambda_j}}, \forall k \in \mathcal{C} \tag{3}$$

中，λ_k 表示观察的类别似然，即对每个类别 $k \in \mathcal{C}$ 的 $\lambda_k = P(S \mid C = k)$ 。这一更新规则借鉴了 Kaplan 关于分类器融合的工作 [?] 。如果其包围框与对应于网格 $\boldsymbol{x}$ 的空间坐标的区域重叠，我们使用得分向量似然 $P(S \mid C = k)$ 来更新 Dirichlet 先验 $\boldsymbol{\beta}^{\boldsymbol{x}}$ 。

D. 中心凹校准

如果我们假设一个检测器确实已经校准良好 [?] ，然后根据贝叶斯法则是 $s_k = P(C = k \mid S) \propto P(S \mid C = k) = \lambda_k$ ，因此实际的原始分数 (1) 可以直接应用于 Kaplan 的规则 (3) 以进行新的更新。

然而，分类器通常没有得到适当的校准，不能反映实际类别的后验概率 (2) ，这与许多分类问题 [?] 中经常假设的相反。为了获得似然值 λ_k ，我们需要从大量数据集中学习适当的传感器模型 $P(S \mid C = k)$ 。

此外，我们必须考虑到这样一个事实：检测器事先已经用常规图像数据集进行了预训练，因此容易受到人工中央凹系统 [?] 引入的周边失真影响。根据人眼的解剖特性 [?] ，视野中某一区域的模糊程度直接与其到中央凹中心的距离相关。因此，我们通过将视场划分为从中央凹中心（即焦点）辐射出的 D 个离散距离层级，以适应中央凹的特征。因此，可以根据各自的边界框所在的距离层级 $d \in \{1, \ldots, D\}$ ，将类别可能性表示为 $P(S \mid C = k, d)$ 。

假设每个 $P(S \mid C = k, d)$ 可以被建模为一个 Dirichlet 分布，我们需要学习 $K \times D$ 个 Dirichlet 分布，每个类别和距离等级各一个，其对应的 Dirichlet 参数为 $\boldsymbol{\alpha}_{k,d} \in \mathbb{R}^K$ 。对于每个类别 k 和距离等级 d 的 Dirichlet 似然 [?] 估计为

$$\mathrm{Dir}(S \mid \boldsymbol{\alpha}_{k,d}) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_{k,d,i}\right)}{\prod_{i=1}^{K} \Gamma(\alpha_{k,d,i})} \prod_{i=1}^{K} s_i^{\alpha_{k,d,i} - 1} \tag{4}$$

，涉及欧拉的 Gamma 函数：$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ 。

客观地讲，每组 Dirichlet 参数是通过拟合多个检测数据进行估计的，这些数据由归一化的得分向量 (1) 组成，根据相应的真实类别 $k \in \mathcal{C}$ 和焦距水平 $d \in \{1, \ldots, D\}$ 进行拟合。鉴于 Dirichlet 分布没有已知的闭式最大似然估计，我们采用由 T. Minka 提出的一种简单迭代方法 [?] ，将从大型数据集中提取的语义数据拟合到 Dirichlet 似然 (4) 上。通过这种技术，我们生成了 $\boldsymbol{\alpha}_{k,d}$ 组数据，借此捕捉由与数据相关的因素（例如光照和遮挡）导致的不确定性，以及由中心凹传感器引入的额外不确定性。

按照图 1 中的虚线路径，我们可以利用新的校准似然 $\mathrm{Dir}(S \mid \boldsymbol{\alpha}_{k,d}) \propto P(S \mid C = k) = \lambda_k$ 来更新信念 $\boldsymbol{\beta}^{\boldsymbol{x}}$ ，同时保持融合规则 (3) 。

E. 主动感知

为了生成预测的注意力图，对于每一个注视点，我们最终调用我们的主动感知机制。本质上，为了构建这些类似显著性的图，我们利用在语义图中累积收集的信息。假设，对于给定的视觉搜索任务，目标是找到目标类 k 的一个实例，我们根据从第 k 个语义图中提取的后验概率 (2) 组装一个概率图。然后，我们贪心地确定 $\boldsymbol{x}^*$ —— 展示最高后验概率的区域，作为下一个最佳注视点：

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmax}}\ P\left(C^{\boldsymbol{x}} = k \mid \boldsymbol{\beta}^{\boldsymbol{x}}\right) \tag{5}$$

我们迭代地应用此决策机制，根据从之前所有注视点 $f_{0:t}$ 收集的信息，选择下一个注视点 $f_{t+1} \equiv \boldsymbol{x}^*$ （以像素为单位），直到最终找到目标对象。除此之外，我们还应

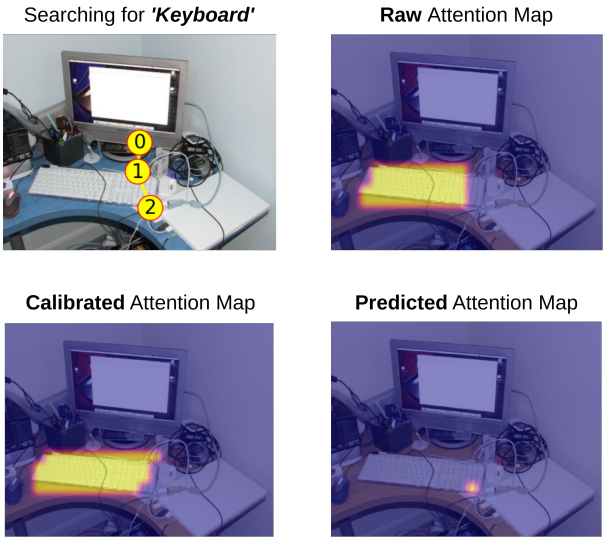

Fig. 2. 在目标存在的视觉搜索设置中，由不同方法生成的注意力（热）图的说明性样例案例。在这个特定的例子中，目标类别是“键盘”，我们展示了经过两次迭代后的地图状态。原始、校准和预测的注意力图分别由 SemBA -FAST Base、Calib 和 Pred 生成和使用。