SPGen：基于无监督域适应的绘画作品随机扫描路径生成

Mohamed Amine Kerkouri[1], Marouane Tliba[2], Aladine Chetouani[2], Alessandro Bruno[3]1F-Initiatives；[2]索邦巴黎北部大学；[2]IULM大学；

arXiv:2602.22049v1 [cs.CV] 25 Feb 2026

摘要

理解人类视觉注意力是保护文化遗产的关键。我们提出了 *SPGen*，这是一种新颖的深度学习模型，可预测扫描路径，即观看者观察绘画作品时的眼球运动序列。

我们的架构采用全卷积神经网络（*FCNN*），结合可微分的注视点选择和可学习的高斯先验，以模拟自然观看偏好。为弥合照片与艺术作品之间的域差异，我们通过梯度反转层进行无监督域适应，使模型能够将从自然场景中学到的知识迁移到绘画作品上。此外，随机噪声采样器用于模拟眼动追踪数据固有的随机性。

大量测试表明，*SPGen* 的表现优于现有方法，可为分析注视行为、促进艺术瑰宝的保护与欣赏提供强有力的工具。

1. 引言

文化遗产涵盖了一个社区创造并代代相传的习俗、传统、文物和艺术表达 [26]。这份集体遗产构成了社区身份认同的重要组成部分，承载着共同价值观与经历的宝库，必须为后代妥善保护。

绘画作品是这份文化遗产的重要组成部分，在人类艺术遗产中占据显著地位。绘画作品是将颜料涂于表面而创作的一种视觉艺术形式，展现了艺术家通过图像传达思想与情感的技巧和创造力 [1]。绘画作品如同视觉叙事，展现了历史事件、社会规范，以及当时的艺术倾向和艺术家本人的创作取向。

绘画作品蕴含着特定时代的美学，以及塑造艺术家视角的文化、政治和哲学基础

[1]https://dictionary.cambridge.org/dictionary/ english/绘画作品

通过珍爱并保护这些艺术瑰宝，我们能够更好地理解过去的细微之处，进而加深对跨越时空的人类创造力之多样性与丰富性的欣赏。

保护并理解这类艺术遗产至关重要，因为它对我们认识人类历史和社会演变有着重大贡献。

对这类媒介的理解与认知，与人类视觉系统（HVS）的认知和感知机制密切相关 [64][10]。人眼拥有卓越的能力，分辨率超过 5 亿像素（MP），因此能够捕捉大量视觉信息。值得注意的是，人类视觉系统具有非凡的能力，人眼接收视觉信息的速率估计约为 10 Gb/s。 [54]。

鉴于需要处理的数据量庞大，而认知资源有限，人类视觉系统（HVS）需要一种高效机制：聚焦于视觉场景中的重要空间区域，筛选出最相关的信息，同时忽略从感知角度而言不那么重要的区域。这种机制称为视觉注意。

具体而言，眼球运动在过滤无关信息方面发挥着至关重要的作用：通过选择性地关注视野中的特定区域并忽略其他区域，实现信息筛选。这种动态机制有效减轻了人类视觉系统的认知负荷，并提高了视觉感知系统的整体效率。

研究发现，在图像观察的早期阶段，颜色、边缘、纹理、强度和对比度等低层视觉特征是吸引眼球运动的主要因素 [63]。这些基本的视觉特征会立即吸引注意力，并引导对视觉场景的初步探索。

相比之下，在视觉体验的后期阶段，语义视觉特征在引导眼球运动方面变得尤为突出

。这些高层次特征包含具有意义的元素，包括面孔、文本、人类和物体，它们在视觉感知和认知中具有相当重要的意义。随着观察者更加投入于图像并理解其内容，注意力会转向这些具有语义意义的方面。

因此，可以根据视觉注意所采用的神经处理通路对其进行分类 [33]。“自下而上注意”主要由外部刺激特征驱动，例如前文提到的低层视觉特征。在这种注意类型中，视觉刺激的内在属性和特征主导着注意机制，这与注意是一种预处理操作的理论相符。自下而上注意以无意识、快速且本能的方式运作，引导我们的注意力转向视觉环境中显著且值得关注的元素。

另一方面，“自上而下注意”受到源自前额叶皮层和长期记忆的高层认知过程影响。这种注意类型由内部因素引导，例如先前的知识、预期和目标，而非仅仅依赖刺激本身的属性。自上而下注意对应于注意是一种后处理操作的理论，因为它涉及有意识且有目的地分配注意力。这种机制具有审慎、缓慢且经过预先思考的特点，使我们能够根据目标和认知策略，有选择地关注视觉场景的特定方面。

上述注意机制可进一步分为两类：“隐蔽”注意和“外显”注意 [69] [1]。隐蔽注意是指在没有伴随任何眼球运动的情况下分配注意力。在这种注意形式下，大脑会优化视觉皮层中的处理活动，优先关注视野中的特定区域。它使观察者无需实际转移注视，也能专注于相关信息。

相比之下，外显注意与眼球运动密切相关。它涉及将中央凹——视网膜上视敏度最高的区域——转向视野中的兴趣区域。因此，外显注意与移动眼睛、使相关区域清晰聚焦的实际行为相一致。通过这种方式，观察者能够从所选区域提取更详细的信息，并以更高的精度进一步处理视觉输入。

隐蔽注意与外显注意的区别，凸显了人类视觉系统动态且适应性强的特性。隐蔽注意使注意焦点能够在视觉场景中快速、灵活地转移，而无需显式的眼球运动；外显注意则借助眼球运动，有策略地仔细检查特定区域。这些互补的注意机制使人类能够高效地探索和理解复杂的视觉环境，为各种认知过程奠定了重要基础。

视觉注意领域的研究主要侧重于对可测量的外显注意进行考察。为研究外显注意，研究人员通常会使用眼动追踪技术，在观看者探索特定视觉刺激时捕捉并记录其眼球运动。眼动追踪仪获取的数据使研究人员能够构建个体与视觉内容交互时的注视轨迹。

注视轨迹由两个主要组成部分构成：注视点和扫视。注视点表示眼睛相对静止并聚焦于视觉场景中特定位置的时刻。在注视点期间，视觉信息被获取并处理。另一方面，扫视是发生在注视点之间的快速眼球运动 [23]。在扫视期间，由于眼睛迅速从一个位置移动到另一个位置，不会发生视觉感知。这种注视轨迹被称为扫描路径，是外显视觉注意的主要主观度量。

多位观察者在观看特定刺激（如图像或视频序列）时所产生的注视点，通常用注视点图来描绘。这些图提供了注视点在刺激上离散分布的可视化表示。

使用与 1° 视角相对应的高斯核对注视点图进行平滑，是获得更平滑的眼球运动表示的常用技术 [38]。这种平滑技术可得到眼球运动的空间分布，有助于降低噪声，并增强对刺激上整体注意模式的可视化。

对注视点图进行平滑处理的一项重要成果是生成显著性图。这些显著性图提供了概率热图，其中每个像素都被赋予一个归一化概率值，用以表示吸引观者注意力的可能性。显著图上概率较高的区域被认为在感知上更为重要，因此更有可能吸引观察者的注意。

视觉注意已成为多个领域（即技术、医学、教育等）以及多个主题中的重要组成部分，例如图像质量评估 [2,3,11–13], 图像和视频压缩 [51],图像描述与说明 [15], 图像搜索与检索 [65], 面向色觉缺陷（CVD）人群的图像增强 [8] [10], 显著性引导的绘画修复 [67] 等等 [20] [61]。

因此，准确地对这一机制的行为进行建模至关重要，也是研究的首要关注点 [59], [60]。

在本文中，我们介绍了一种新颖且高效的向量到序列建模方法，专门用于将图像转换为扫描路径。我们的方法

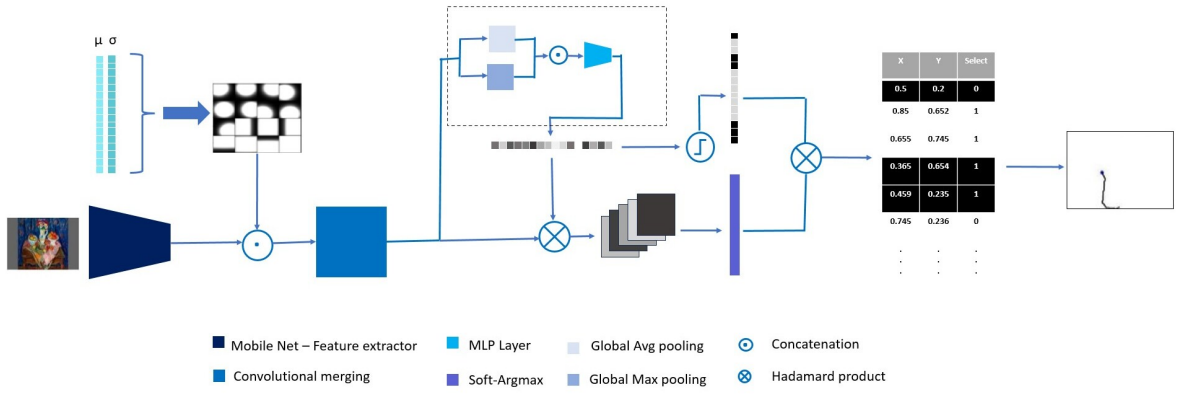


图1：所提方法的总体架构。

利用全卷积神经网络来完成这项任务。

我们在研究中要解决的一个挑战是，来自不同域的图像存在各种偏差，例如自然场景、艺术作品和合成数据。这些域特定偏差与视觉感知任务相关的主观偏差相结合，会形成带有噪声的分布，妨碍模型准确捕捉图像特有且与任务相关的特征。

我们探索了两种领域自适应方法来解决这一问题。第一种方法着眼于制定下游任务先验偏差，第二种方法则侧重于泛化特征提取，也称为域表征对齐 [27]。这些方法应用于多种数据分布域，以提高绘画数据集上的预测准确率。

神经网络表现出确定性行为，这使得在其中引入噪声颇具挑战。鉴于神经网络的确定性，在其中引入噪声是一项关键策略。对于给定输入，神经网络通常会生成固定输出，这可能会限制生成结果的多样性，尤其是在处理注视扫描路径这类随机且主观的现象时。因此，考虑到注视扫描路径的随机性和主观性，我们在表征的潜在空间中引入了一个随机噪声采样器。这有助于有效控制模型的可变性，而这种可变性由温度参数调节。温度参数通过捕捉人类注视行为内在的不确定性和主观性，控制噪声的强度。

本工作的主要贡献总结如下：

- 提出一种高效的扫描路径预测深度学习模型。
- 引入一个选择模块，使网络能够生成可变长度扫描路径。
- 在模型中引入随机机制，为相同刺激生成多条扫描路径。
- 纳入一个任务领域特定模块，以考虑先验知识。
- 通过无监督对抗领域自适应，增强我们的模型对绘画作品图像的泛化能力。这一方法能够对齐特征分布，从而无需标注的绘画图像数据即可实现准确预测。
- 在自然场景和绘画图像数据集上开展全面的定性与定量实验，以评估所提出方法的有效性。

本文其余部分安排如下：
第2 节概述了视觉注意力建模相关的既有研究，涵盖显著图和扫描路径预测。第3节介绍了我们提出的模型架构各组成部分、训练方案，以及如何利用无监督领域自适应实现泛化。第4节详细展示了模型性能的结果与评估，并对其各组成部分进行了消融实验。第5节总结了本研究的主要发现。

2. 相关研究

本节全面概述显著性预测和扫描路径检测的相关研究，包括传统模型和基于深度学习的模型。

2.1. 显著性预测

显著性预测历史悠久，始于 [38], 开创性研究，该研究由 [28] 使用基于低层特征（如

颜色、强度和方向。其他模型，例如 [9], 提出的模型，则基于提取关键点，在多尺度分析中进行显著性预测。

过去十年深度学习的成功推动了各种用于显著性预测的深度架构模型的发展。例如， [39]使用MIT数据集 [32] 训练深度网络，并在后续版本中进一步改进了结果 [40, 41]。类似地， [29] 使用多尺度架构和大型数据集进行训练。

在 [19], 中，作者提出了一种适用于静态图像和视频的统一显著性预测模型（Unisal）。他们利用高斯先验图对偏差进行建模，从而获得了更好的预测结果。

显著性预测领域也采用了基于图的方法进行研究。例如，GBVS（基于图的视觉显著性） [25], 由 [**?**], 提出，它依赖于在不同输入图上定义的图和马尔可夫链。与此同时， [7] 提出了一种基于信息最大化的模型，该模型使用香农自信息度量。

一些研究在谱域中探讨了显著性预测，例如 [52] 和 [24]。相比之下，Bruno等人 [9] 在近似感知均匀的颜色空间中使用局部关键点进行显著性预测。

深度学习范式的兴起对显著性预测产生了积极影响。例如， [50]提出了一种利用深层和浅层神经网络进行显著性预测的方法。DeepGaze I [39] 及其改进版本 [40]利用MIT数据集和VGG-19网络基础，实现了准确的显著图预测。

其他用于显著性预测的深度学习方法依赖于网络不同层级的多层级特征，如 [14]和 [29]所示。此外，Sal-Gan [49] 采用编码器-解码器架构，结合生成式建模与对抗学习。

[16] 提出了一种将卷积神经网络与循环长短期记忆（LSTM）相结合的模型，以捕捉注意力机制。Unisal [19] 采用高斯先验图来对图像和视频中的偏差进行建模。

[47] 将卷积神经网络与Transformer架构相结合，有助于捕捉图像区域之间的长距离依赖，从而进一步提升预测效果。

Tliba 等人 [59] 采用类似于 [**?**] 的 U-Net 架构，并在长跳跃连接上使用自注意力头。该模型尝试融合早期阶段提取的低层特征和后期阶段提取的高层特征。使用自注意力有助于进一步提升模型的表征能力。

2.2. 扫描路径预测

扫描路径预测是一个相对较新的研究领域，但近年来受到了广泛关注。

[28]提出的赢家通吃（WTA）模块是最早用于预测扫描路径的方法之一。该方法从显著性图峰值中提取注视点，并抑制这些点周围的区域。

Le Meur 等人 [44], 提出了一种生成扫描路径的随机方法。具体而言，他们预测显著性图，并对若干偏差（例如扫视幅度和方向）的概率进行建模。

[68] 将显著性图视为二维引力场，它会影响代表注视的质量块的运动轨迹。该引力场会影响质量块的速度和运动轨迹，而这对应于注视运动。

另一种方法由 [57], 提出，该方法利用来自卷积神经网络的高层特征和记忆偏差（包括短期记忆和长期记忆）来预测扫描路径。

[5]提出了 Saltinet，这是一种深度学习模型，可从编码器-解码器网络生成的静态显著性体中推断扫视。推理采用不同的采样策略，从而生成扫描路径。随后，同一批作者提出了 PathGan [4],，该模型采用 LSTM 网络和条件 GAN 架构，以更高的准确率预测扫描路径。

DCSM（深度卷积扫视模型） [6] 预测中央凹显著性图和时间持续时长，同时对返回抑制（IoR）进行建模，以反映扫描路径的动态时间特征。

SALYPATH 模型 [37], 提出了一种端到端模型，可同时预测图像的扫描路径和显著性图。该方法利用这两项任务之间的相关性来提升性能。随后，同一方法被推广到 360° 图像 [**?**]。

[60] 的研究采用轻量级架构，在中央偏置理论的基础上进行扩展，并使用受 Point-Net 启发的一维卷积来预测合适的扫描路径。作者采用自监督训练方法，将模型扩展至绘画数据集，并首次推出 AVAtt（艺术视觉注意力）数据集的初步版本。

[34], 的研究使用无监督领域自适应 [62] [22], 将一个基于自然场景图像训练的定制模型中的知识迁移到绘画作品上。这项工作是该模型的扩展。

作者还在 [35] 中使用对抗训练来提升模型性能，并帮助捕捉观察者间的模式，从而实现一致的预测。

在 [56] 中，作者受到人类视觉系统生物学模型的启发，采用了中央凹预测模型。

在 [17], 中，作者引入了受语义内容和注视历史影响的动态优先级图。该模型利用卷积神经网络、卷积长短期记忆网络和混合密度网络生成模拟人类扫描路径。

在 [66]《显著性与扫描路径统一模型》

（UMSS）是一种能够同时预测信息可视化中的视觉显著性和扫描路径的方法。UMSS 分析各个可视化元素的注视行为，并呈现多持续时间的元素级显著性图。随后，它根据这些图以概率方式采样生成扫描路径。

[58] ScanDMM 是一种深度马尔可夫模型（DMM）架构，用于预测 360度图像中的扫描路径。该模型考虑了时间依赖性，并引入语义引导的转移函数，以学习全景图像中注意力景观的非线性动力学。

总而言之，显著性预测历史悠久，传统模型和基于深度学习的模型都取得了令人瞩目的成果。扫描路径预测是一个较新的研究领域，近年来取得了显著进展，相关方法利用随机模型、引力场类比、高层特征和动态时间特性来实现准确预测。结合显著性预测模型与扫描路径预测模型，有助于更全面地理解视觉注意机制和图像理解。

3. 提出的方法

3.1. 架构组成

本文提出了一种用于扫描路径预测的新型框架，并展示了无监督域适应在这一任务上的有效性。所提出模型的整体架构如图1所示。模型的关键组件和步骤如下：

1. **特征提取器**（图像编码器）：MobileNet 用作特征提取器，负责高效地从输入图像中捕获相关模式与特征。

2. **可学习域先验偏置图**：可学习域先验偏置图与特征提取器的输出拼接，以帮助模型在无监督领域自适应过程中理解特定于域的特征。

3. **合并卷积**：合并后的特征图（来自特征提取器与偏置图）经过一系列卷积处理，以提取与扫描路径预测相关的特定特征。

4. **Soft-ArgMax**：加权特征图通过 Soft-Arg Max 函数处理，以生成预测注视点的连续坐标。

5. **注视点选择器**：该组件利用全局池化与全连接层对概率向量进行离散化，并生成与相关预测注视点对应的二值掩码。

6. **随机生成**：通过噪声采样并使用温度参数进行调制来引入随机性，从而为扫描路径的生成增添随机性。

通过结合这些步骤和组件，所提出的模型能够有效预测扫描路径，并通过无监督域适应适应新域。随机生成在预测中增添了随机性，体现了人类注视模式的主观性。

3.1.1 特征提取器

注视预测中的逐像素计算涉及逐一分析和处理输入图像中的每个像素。由于图像具有高维复杂性和空间相关性，这类计算的计算量可能很大。为应对这种复杂性并减轻计算负担，我们使用卷积神经网络（卷积神经网络）对输入图像进行建模，并提取有意义的特征。

卷积神经网络 [21] [45]是一类非常适合图像相关任务的神经网络，因为它们能够自动学习图像数据的层次化表示。借助卷积神经网络，我们可以从输入图像中捕获相关模式和特征，这对于准确进行注视预测至关重要。

在我们提出的方法中，我们具体使用 MobileNet[55]作为特征提取器。MobileNet 是一种轻量级卷积神经网络架构，专为移动和嵌入式应用而设计。它采用深度可分离卷积和逐点卷积，大幅减少网络中的参数数量和计算量。因此，这种架构效率更高，也更易于训练，尤其适用于我们当前数据量有限的情况。

通过采用 MobileNet 作为特征提取器，我们的整体架构依然轻量，同时能够从输入图像中提取关键特征。选择这一网络有助于提升注视预测模型在给定数据集和计算资源条件下的效率与效果。

3.1.2 优先级图与注意偏差

图像中的显著性往往呈现一种称为中心偏差的现象，即与边缘和角落相比，观察者倾向于将更多注视集中在图像的中央区域。这种所谓的中心偏差现象在眼动追踪实验中很常见。人们认为，这是人类视觉系统倾向于关注相关语义信息所致，而这些信息通常集中在图像的中央区域。这与视网膜上感光受体的解剖分布有关 [30]。

为了建模中心偏差现象，一种常用方法是使用二维高斯函数来表示它。高斯函数是一种概率分布，它为更靠近中心的点分配更高的概率，为距离更远的点分配更低的概率，远离

中心。二维高斯函数的数学形式由公式1：

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right) \tag{1}$$

在此公式中，x和 y 是图上某一点的坐标，而 (μ_x, μ_y) 和 (σ_x,σ_y) 分别是该分布的平均值和标准差。这些参数 (μ_x, μ_y) 控制高斯函数峰值的位置，而 (σ_x,σ_y) 则决定分布的扩散程度或宽度。

在我们提出的模型中，我们利用这些高斯函数对扫描路径预测中的中心偏差进行建模。然而，我们的模型并非使用单个高斯函数来表示中心偏差，而是学习多个图，这些图都遵循相同的高斯规律，但具有不同的平均值和标准差。这些图称为“注意力偏置”或“优先级图”。

将这些注意力偏置纳入模型后，我们便能让特征提取器专注于对刺激图像的特定特征进行建模，而不受显著性任务相关偏差的影响。换言之，注意力偏置有助于模型忽略眼动追踪数据中普遍存在的中心偏差现象，转而专注于学习特定刺激所对应的显著性模式。

整合注意力偏置对于实现准确且相关的扫描路径预测至关重要，因为这有助于模型更好地理解和解读图像中的显著性模式。通过从数据中学习注意力偏置，模型能够有效处理中心偏差现象，并生成与人类眼动行为更加吻合的扫描路径预测。

3.1.3 扫描路径预测

扫描路径预测的解码过程由多个部分组成，分别是合并模块、注视点选择网络和 Soft-ArgMax 函数。

合并模块是一个由 8 个 (3×3) 卷积层组成的网络，每层均由线性整流单元（ReLU）函数激活。该模块以 MobileNet 特征提取器提取的特征图为输入，并将其与前文所述的可学习先验图（注意力偏置）拼接。合并网络将 MobileNet 提取的刺激特定特征与先验图相结合，同时逐步将特征图数量减少至 20。

通道选择网络会生成一个激活概率向量，用于对合并模块生成的特征图进行加权。通道选择网络负责预测注视点使用 Soft-ArgMax 函数（SAM）确定坐标 [16]，与 [37]中的方法类似。

Soft-ArgMax 函数根据模型输出的特征图预测注视点的坐标。其数学描述如下：

$$SAM(x) = \sum_{i=0}^{W}\sum_{j=0}^{H}\frac{e^{\beta x_{i,j}}}{\sum_{i'=0}^{W}\sum_{j'=0}^{H}e^{\beta x_{i',j'}}}\left(\frac{i}{W}, \frac{j}{H}\right)^T \tag{2}$$

在此公式中，x 是输入特征图，(i,j) 和 (i',j') 遍历像素坐标。 W 和 H 分别表示特征图的高度和宽度。参数 β 用于调整 softmax 输出的分布。

注视点选择器是另一个关键组件，它会生成一个二值掩码向量，用于选择合适的注视点并将其纳入预测扫描路径。这使得生成可变长度扫描路径成为可能。注视点选择器的配置方式是：将合并模块输出的特征图展平后，并行输入全局最大池化层和平均池化层。得到的向量作为浅层多层感知机（MLP）网络的输入。MLP 两个分支的输出会拼接后传入第三个 MLP 层，该层的最后一层由 Sigmoid 函数激活，从而输出一个取值范围为 $[0,1]$ 的连续值向量。然后根据该向量的平均值对其进行动态二值化，以生成二值掩码向量，如下式所示：

$$mask_b = v > mean(v) \tag{3}$$

with

$$v = MLP_3(cat(MLP_1(GMaxPool(flat(d))), MLP_2(GAvgPool(flat(d))))) \tag{4}$$

其中 $mask_b$ 是生成的二值掩码向量。MLP_1, MLP_2 和 MLP_3 是浅层多层感知机网络，cat 表示向量拼接， $GMaxPool$ 表示全局最大池化操作，$GAvgPool$ 表示全局平均池化操作， $flat$ 表示展平函数， d 是从合并层获得的输出特征图。

由于二值化函数的不可微性，网络的计算图被切断，从而阻断了梯度的流动。为解决这一问题，将从多层感知机获得的向量与特征图相乘。这确保了二值化函数的不可微性得到有效处理。换言之，由于二值化函数不可微，我们需要找到另一种方式来允许梯度流动。因此，我们将注视点选择网络的结果与SAM函数的输入相乘。这一步骤使我们能够保留必要的梯度信息，并确保训练与优化顺利进行。

为解决二值化函数的不可微性，网络的计算图被切断，并将从多层感知机获得的向量与特征图相乘。多层感知机最后一层的Sigmoid激活确保运算保持在与网络输出相同的范围内，从而避免反向传播过程中的梯度爆炸问题。该注视点选择网络在生成准确的扫描路径预测中发挥着关键作用，因为它能基于特征图有效选择相关注视点，从而生成与人眼运动行为高度相似的扫描路径。

3.2. 训练与损失

训练过程使用了 Salicon 数据集中的 9000 张图像 [29], 另有 1000 张图像留作验证。

为了训练模型，我们采用了以下损失函数:

$$L(y,\hat{y}) = BCE(y,\hat{y}) + 0.001 \times \sqrt{\text{len}(y)^2 - \text{len}(\hat{y})^2} \quad (5)$$

在该公式中， y 表示预测的扫描路径向量，$\hat{}y$ 表示真实标注的扫描路径向量， BCE 表示二元交叉熵函数。

该损失函数旨在捕捉扫描路径预测的两个方面。首先，它使用二元交叉熵损失惩罚模型与真实标注之间的不匹配。其次，它引入了一个正则化项，该项基于预测扫描路径与真实标注扫描路径之间的长度差异。

训练时使用 Adam优化器，学习率为 5×10^{-5}。我们对模型进行了 70 个周期的训练，并使用 Xavier 初始化器初始化所有权重。这样的训练设置使模型能够有效地学习并调整参数，从而预测准确且具有代表性的扫描路径。训练期间的验证过程确保能够监控并优化模型性能。

3.3. 领域自适应

如第4节所示，我们的模型在自然场景图像上的初始训练取得了良好的效果。然而，当我们在其他源域的图像（如绘画作品）上测试同一模型时，效果却不尽如人意。为了充分发挥模型的能力，我们需要使其适应新的源域。

领域自适应是指在源域（例如自然场景图像）上训练模型，然后对其进行自适应，使其在目标域（例如绘画作品）上表现良好。关键挑战在于最小化两个域分布之间的差异，从而使模型能够有效泛化到目标域。为此，我们采用一种自适应方法，该方法提出于 [22]。

为了实现领域自适应，我们希望特征提取器能够识别两个域之间共享且有用的信息，同时丢弃与预测任务无关的域特定特征和噪声。为此，我们在网络中引入一个小分支，用于将图像分类为属于自然场景域或绘画领域。该分支以伪标签方式添加，这意味着我们没有显式的域标签，而是将源域和目标域视为不同的类别。

为了促使特征提取器在两个域之间找到统一的表征空间，我们在域分类器分支的起始处使用了梯度反转层（GRL）。GRL 会在反向传播过程中反转梯度流的符号。在前向传播过程中，它不会产生任何影响（式 6），但在反向传播过程中，它会将梯度乘以 -1（式 7）。

$$GRL(x) = x \quad (6)$$

$$GRL(\frac{\partial L}{\partial x}) = -\frac{\partial L}{\partial x} \quad (7)$$

其中， x 是该层的输入， $\frac{\partial L}{\partial x}$ 表示通过网络进行反向传播时的域损失 L_d 的梯度。

如图**??**所示，当域分类器学习区分来自两个域的图像时，GRL中梯度符号的反转会促使特征提取器寻找共享表征空间，从而最小化两个域分布之间的距离。

本质上，领域自适应过程迫使特征提取器关注两个域的共同特征，同时忽略域特定变化，从而有效对齐两个域的分布。

统一表征空间可以概念化为两个分布（自然场景图像和绘画作品）的并集减去每个域的特定噪声分布:

$$\mathcal{D} = \{\mathcal{D}_p \cup \mathcal{D}_n\} \setminus \{\mathcal{N}_p \cup \mathcal{N}_n\} \quad (8)$$

其中:

- $\mathcal{D}_n$ 和 $\mathcal{D}_p$ 分别是自然场景图像和绘画作品的分布。
- $\mathcal{N}_n$ 和 $\mathcal{N}_p$ 分别是自然场景域和绘画领域特有的噪声分布。

对于无监督域适应训练，我们混合使用了 2000 张 Salicon 图像 [29] 和从互联网上抓取的 2000 张未标注绘画作品图像。这使模型能够适应绘画领域的特征，并在自然场景图像和绘画作品图像上均表现良好。

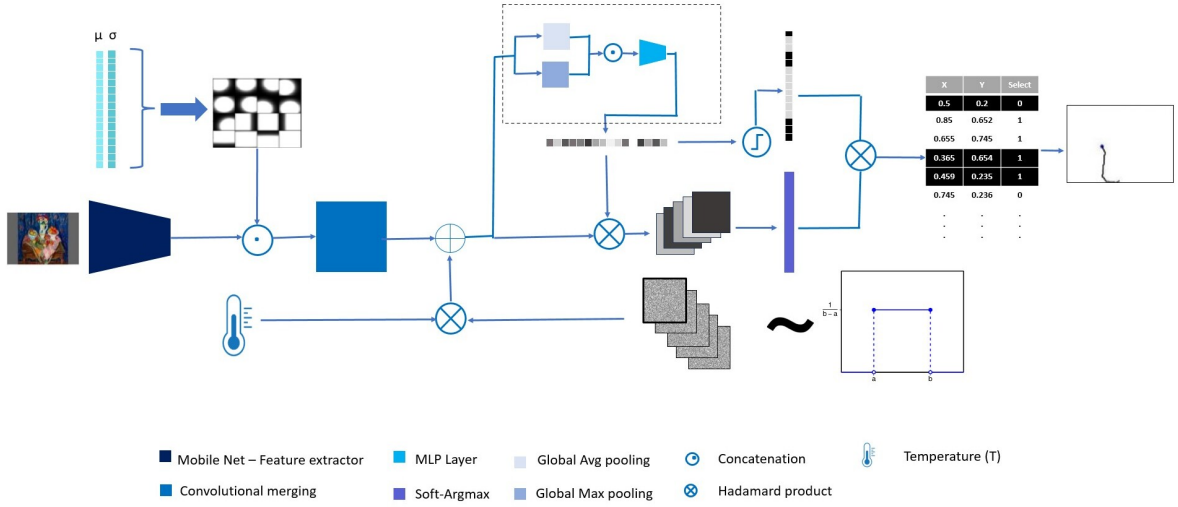


图2：所提出方法的总体架构。

3.4. 随机推断

视觉注意是一种复杂且主观的现象，受到多种生理、心理、认知、环境及情境因素的影响。因此，我们观察到，不同受试者之间，甚至同一个人在不同情境或不同时期，其注视模式和扫描路径都存在显著差异。这些差异在眼动追踪实验中无法得到完全控制或解释，这表明扫描路径预测是一项随机生成任务。

为了在模型的预测中引入随机性，我们按照图 2 所示修改了网络架构。我们引入一个新的元素，即随机潜在向量 $L \sim U(\text{待填充})$，其中$L \in \mathbb{R}^{w \times h \times c}$。该随机向量捕捉了扫描路径生成过程的随机特性。随后，我们通过变量 T 调节随机性，并将其称为温度。温度参数控制生成的扫描路径的随机性程度。

我们可以如下对生成的扫描路径进行建模：

$$y = \text{decoder}(\text{encoder}(x) + L \times T) \tag{9}$$

其中：
x 是输入图像。encoder 是模型中将输入图像转换为潜在表示（即特征提取器）的编码器部分。decoder 是模型中的解码器部分，它根据潜在表示和经过调制的随机性 $L \times T$生成预测扫描路径。通过引入随机潜在向量，并利用温度调节其影响，我们可以在扫描路径生成过程中加入随机性。这使模型能够为同一输入图像生成多样且各不相同的扫描路径，反映人类注视行为中固有的变化性。温度参数控制模型预测中确定性行为与随机性行为之间的平衡，使其能够灵活生成随机程度不同的扫描路径。

4. 实验方案

4.1. 数据集

在本研究中，我们采用了两个广受认可的自然场景数据集，即 Salicon [29] 和 MIT1003 [32]。此外，我们还使用了两个绘画数据集，即 Le Meur 绘画数据集 [43] 和 Art-Visual-Attention（AVAtt）数据集 [60] [36]。这些数据集为我们的实验评估奠定了基础，使我们能够评估模型在多种视觉领域（包括自然场景和绘画作品）中的性能。

4.1.1 Salicon 数据集

该**Salicon** 数据集是一个从 MS-COCO（上下文中的常见物体）数据集中提取的大规模且极具价值的资源 [46]。它包含种类丰富的图像，每张图像都附有相应的显著图和一组记录的扫描路径（≈ 50）。该数据集是评估和改进视觉注意力建模方法的重要基准。

在实验中，我们将 Salicon 数据集划分为三个子集：包含 9,000 张图像的训练集、包含 1,000 张图像的验证集，以及包含 5,000 张图像的测试集。训练集和验证集用于训练我们提出的模型，而测试集则在评估模型性能方面发挥了关键作用，

并将其与该领域其他最先进模型进行比较。

在整个工作中，我们通过对 Salicon 数据集中约 250,000 条可用扫描路径进行测试，广泛评估了我们的模型 [29]。如此大量的扫描路径确保了实证结果的可靠性与稳健性，使我们能够全面考察模型性能的各个方面。广泛的评估过程，结合大规模且多样化的数据集，增强了研究结果的有效性与可靠性，使我们能够全面理解所提出模型的有效性与能力。

4.1.2 MIT1003数据集

MIT1003数据集被广泛认为是静态图像显著性领域最著名的数据集之一。它常与**MIT300** 数据集 [31], 一同使用。MIT1003包含1,003张自然场景图像，这些图像在不同条件下描绘了各种物体、人物和场景。数据集中的每张图像都配有相应的显著图，以及由15名不同观察者记录的一组扫描路径，共计15,045条可用于测试的扫描路径。

在我们的研究中，我们主要使用完整的MIT1003数据集来评估我们的模型在不同数据集上的泛化能力，具体用于跨数据集评估。为此，我们避免在该数据集上对我们的模型进行微调，而仅专注于评估其在先前未见数据上的性能。这种方法使我们能够衡量模型将其预测泛化到训练数据之外的能力，并评估其对MIT1003数据集中遇到的新颖视觉场景和多样条件的适应能力。

通过在未经任何微调的情况下在MIT1003数据集上进行评估，我们可以有把握地分析我们的模型泛化能力的稳健性与有效性，从而深入了解其在先前未见数据上做出准确预测的能力。该数据集中大量的扫描路径以及图像的多样性有助于全面评估我们的模型性能，并展示其在现实应用中的潜力。

4.1.3 Le Meur绘画作品数据集

Le Meur**绘画作品**数据集 [43] 是专为绘画作品领域的显著性建模而设计的宝贵资源。该数据集包含 150 幅绘画作品图像，每幅作品都附有相应的显著图。该数据集是一个关键的评估基准，用于评估我们的模型将从自然场景图像中学到的知识迁移到绘画作品这一独特领域的能力。此外，它还使我们能够在这一独特的数据集上测试域适应技术的有效性。

尽管 **Le Meur 绘画作品** 数据集不提供观察者的扫描路径，但它仍是少数公开可用的绘画作品视觉注意力数据集之一。尽管缺少扫描路径，该数据集仍是评估我们的模型在绘画领域中的泛化能力的重要资源。通过应用我们的领域自适应技术，我们可以探究模型对绘画作品独特特征的适应能力，以及它能否在这一专业领域中准确预测显著性。

Le Meur 绘画作品 数据集在将我们的研究范围拓展至自然场景图像之外方面发挥着关键作用，并展示了我们的模型处理多样化视觉领域的灵活性。借助该数据集，我们能够深入了解模型泛化其知识并在绘画作品这一独特语境中有效预测显著性的能力，从而加深对各种艺术和视觉内容场景中视觉注意的理解。

4.1.4 Art-Visual-Attetnion 数据集

Art-Visual-Attention 数据集（AVAtt） [60] [36] 包含 164 幅图像，每幅图像均附有对应的显著图以及 15 条观察者扫描路径。作为目前唯一一个为每幅图像提供独立扫描路径标注的数据集，AVAtt 是研究艺术作品语境下视觉注意的重要资源。

尽管 AVAtt 数据集目前仍在持续扩充，尚未公开分享，但它在我们的研究中发挥着至关重要的作用。我们利用该数据集评估模型在新绘画图像上的泛化能力，具体而言，就是评估我们的模型适应多种艺术风格和艺术运动的能力。该数据集包含来自 16 种不同艺术运动的丰富多样的图像，描绘了各种各样的主题。

通过在 AVAtt 数据集上测试我们的模型，我们深入了解了模型将其在其他数据集上训练所获得的知识迁移到艺术作品独特特征与风格上的能力。数据集中还为每张图像提供了单独的扫描路径标注，进一步提升了其价值，使我们能够更细致地分析模型的表现，并了解模型如何预测具有各种艺术影响的绘画作品中的视觉注意。

总的来说，AVAtt 数据集是探索我们模型对新绘画图像的泛化能力和适应性的关键资源，可为艺术分析、视觉内容设计以及其他需要理解艺术作品中人类视觉注意力的领域提供有关现实应用的宝贵信息。

4.2. 评估指标

在评估中，我们采用了第4.1 节所述的同一组图像和数据划分，对我们的模型和对比方法进行测试。为了评估各个模型的表现，我们使用三个评估指标，将它们的结果与真实标注进行比较： $MultiMatch$ [18], NSS[53], 和 $Congruency$ [42]。

4.2.1 Multimatch 指标

MultiMatch （Jarodzka）由 Jarodzka 等人提出[18]。该指标以多维向量的形式度量扫描路径之间的相似性，并基于五种不同特征对它们进行比较：

1. **形状**：它量化以向量表示的扫描路径之间的形状相似性，记为 $u - v$。该计算捕捉预测扫描路径与真实扫描路径之间的结构差异，从而能够对其空间构型进行全面比较。

2. **方向**：它评估扫描路径中扫视之间的角度差异。该评估确保预测的扫视方向变化与人类注视行为中观测到的分布相一致，从而验证模型生成真实、自然注视转移的能力。

3. **长度**：它评估扫视之间的长度差异。该评估量化模型预测扫描路径与观测扫描路径之间注视转移幅度的变化，从而为模型在复现人类眼球运动空间特征方面的准确性提供有价值的见解。

4. **位置**：计算扫描路径中各注视点位置之间的差异。通过这一计算，可以全面评估模型预测的注视点与人类观察者实际注视位置的吻合程度。量化位置差异后，该特征可衡量模型捕捉观测扫描路径中视觉注意空间分布的准确性。

5. **持续时间**：评估模型预测扫描路径与观测扫描路径之间的扫视持续时间差异。通过这一评估，可以全面分析模型捕捉视觉注意时间特征以及人类眼球运动中观察到的注视转移时机的能力。然而，由于我们的模型不预测时间维度，因此在评估中排除了这一组成部分。因此，我们仅关注**MultiMatch** 指标的其他特征，以评估模型在视觉探索过程中复现人类注视空间转移的性能。

为使扫描路径比较切实可行，MultiMatch 指标包含三个关键步骤。首先，通过合并较小的、局限于局部的扫视和注视点来简化扫描路径。这一关键阶段确保细微的差异不会过度影响整体相似度结果。其次，对简化后的扫描路径进行时间对齐，从而为后续比较建立扫视对之间的对应关系。Dijkstra 算法依据潜在扫视配对之间的差异最小化代价，从而有效确定最优对齐。最后，对已对齐的扫描路径进行详细比较，并综合考虑五个维度，包括上文所述的扫视幅度、方向和注视持续时间。

每项特征的取值范围均为 0 到 1，数值越低表示扫描路径之间的相似度越高。

为了获得一个代表整体相似度的总分，我们对前四项特征（即形状、方向、长度和位置）取算术平均值；由于我们的模型不预测注视时间戳，因此不包括持续时间。整体分数记为*MM* 分数，范围为 0 到 1，其中 1 表示相似度高，0 表示相似度低。

MultiMatch 指标全面评估扫描路径的相似度，使我们能够评估模型捕捉人类视觉注意模式基本特征的性能，并将其与真实扫描路径进行比较。

4.2.2 NSS 指标

NSS （归一化扫描路径显著性） [53] 是一种用于比较显著图与预测扫描路径的混合评估指标。该过程包括根据预测扫描路径构建二值注视图。随后，将该二值图与归一化显著性图逐像素相乘；归一化显著性图的平均值为零，标准差为一。将相乘结果取算术平均值，即得到 NSS 分数，该分数表示扫描路径中注视点的平均显著性值。

NSS 分数由以下公式给出：

$$NSS = \frac{1}{N} \sum (P \odot Q_b) \tag{10}$$

其中：

- **N** 是扫描路径中的注视点总数。
- $\mathbf{Q_b}$ 是由扫描路径导出的二值注视图。
- **S**是原始显著图。
- **P** 是归一化显著性图，计算方式为$\frac{S-\mu(S)}{\sigma(S)}$ 。
- $\mu(S)$ 和 $\sigma(S)$ 分别表示显著图的平均值和标准差。

NSS 分数用于衡量模型预测扫描路径与显著图中总体显著性分布的对应程度。较高的 NSS 分数表明预测注视点与显著图中显著性较高区域的对齐更好，从而体现模型在复现人类观察者所表现出的由显著性驱动的注意模式方面的准确性。

4.2.3 一致性指标

该一**致性**指标是一种混合评估方法，通过评估预测扫描路径与真实标注之间的空间重叠和相似性来衡量两者的一致程度。较高的一**致性** 分数表示与观察到的注视模式更加吻合，说明模型能够准确复现人类的注意焦点。

要计算一**致性** 分数，请执行以下步骤：

$$Congruency = \frac{\sum(P_{otsu} \odot Q_b)}{\sum Q_b} \tag{11}$$

其中

- $\mathbf{Q_b}$ 是由扫描路径或扫描路径集合生成的二值注视图。
- **S** 是原始显著图。
- **P** 是归一化显著性图，计算方式为$\frac{S-\mu(S)}{\sigma(S)}$ 。
- $\boldsymbol{P_{otsu}}$ 是通过应用大津阈值算法计算得到的二值化显著图 [48]。

一致性评分能够揭示模型预测的注视点与刺激中高显著性区域的吻合程度。较高的一致性评分表明，模型的注视位置与人类观察者识别出的显著区域更加一致。该指标特别适用于评估模型对刺激中视觉显著区域注意焦点的捕捉能力，反映模型再现类人注意模式的能力。

图 3总结了一**致性**指标的主要步骤。首先，根据真实扫描路径构建注视图，并通过高斯核卷积进行处理，为图像生成显著图。随后，使用大津法对该显著图进行动态阈值处理 [48] ，以区分显著区域和非显著区域。一**致性**的计算方式为：二值图中显著区域内的注视点数量与扫描路径中注视点总数之比。

与 NSS 不同，一**致性** 指标不会考虑每个注视点的显著性值，而是在评估过程中将所有显著点一视同仁。

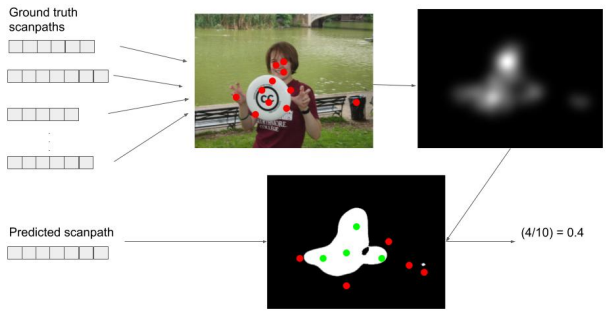


图 3：计算预测扫描路径的一致性。

4.3. 实验方案概述

为了验证我们的模型，我们使用第 4.1 节中的数据集以及第 4.2 节所述的指标。

为展示我们针对自然场景图像的判别模型（第 4.4 节）的有效性，我们使用了 Salicon 数据集中的 5000 幅图像 [29] 及其对应的 $\approx 250,000$ 条对应扫描路径。我们还使用了 MIT1003 [32] 中的 1003 幅图像及其对应的 15045 条扫描路径。

这使得 Salicon 数据集需要比较的理论配对数量达到 $(\approx 3,750,000)$ 组。由于计算资源有限，且 MultiMatch 指标 [18],的计算开销较大，同时没有可利用硬件加速的公开实现。

我们选择仅在 MIT1003数据集 [32] 上评估生成模型，详见第 4.6 节。这需要进行 $(225,675)$ 次比较以完成评估。作出这一决定也考虑到，MIT1003 是真正的眼动追踪数据集，而 Salicon 使用的是鼠标移动数据。

我们使用作者通过定性方式选取的五个温度值对生成模型进行了评估（第4.6节）；我们采用正态分布和均匀分布来生成扫描路径噪声。

在4.5节中，我们对判定架构中使用的各个组件进行了消融研究。

我们测试了领域自适应方法在判定方法上的有效性，随后评估了受益于领域自适应的生成架构。

最后，在4.8节中，我们展示了本方法的一些定性结果。

4.4. 判定网络的定量评估

表 1 展示了在 Salicon 数据集上进行扫描路径预测的结果。最高分数以粗体标出。我们的模型在 $Shape$、 $Length$、

模型	形状	方向	长度	位置	MM 分数	NSS	一致性
PathGAN [4]	0.9608	0.5698	0.9530	0.8172	0.8252	-0.2904	0.0825
Le Meur [44]	0.9505	0.6231	0.9488	0.8605	0.8457	0.8780	0.4784
G-Eymol [68]	0.9338	0.6271	0.9521	0.8967	0,8524	0.8727	0.3449
SALYPATH [37]	0.9659	**0.6275**	0.9521	0.8965	0,8605	0.3472	0.4572
我们的模型	**0.9702**	0.6173	**0.9587**	**0.8968**	**0,8607**	**1.0140**	**0.5170**

表1：Salicon 数据集上的扫描路径预测结果

模型	形状	方向	长度	位置	MM 分数	NSS	一致性
PathGan [4]	0.9237	0.5630	0.8929	0.8124	0.7561	-0.2750	0.0209
DCSM (VGG) [6]	0.8720	0.6420	0.8730	0.8160	0,8007	-	-
DCSM（ResNet）[6]	0.8780	0.5890	0.8580	0.8220	0,7868	-	-
Le Meur [44]	0.9241	0.6378	**0.9171**	0.7749	0,8135	0.8508	0.1974
G-Eymol [68]	0.8885	0.5954	0.8580	0.7800	0,7805	0.8700	0.1105
SALYPATH [37]	0.9363	**0.6507**	0.9046	0.7983	0,8225	0.1595	0.0916
我们的模型	**0.9392**	0.6152	0.9100	**0.8537**	**0.82952**	**0.8888**	**0.2114**

表 2：MIT1003 上的扫描路径预测结果。

和 $Position$ 组件上取得了可接受的结果，同时在 $Direction$ 组件上也表现良好。在代表四个组件平均分的 *MM* 分数上，我们的模型略微超过了排名第二的模型。

我们的模型在 $MultiMatch$ 指标上的性能与最先进模型相当，甚至在某些评价标准上略优于 SALYPATH。在 NSS 指标方面，我们的模型取得了最佳结果，超过了使用显著图生成扫描路径的 Le Meur 和 G-Eymol。不过，在 $Congruency$ 指标上，我们的模型表现不及 Le Meur 和 SALYPATH。这可能是因为 Le Meur 高度依赖显著图来生成扫描路径，而 SALYPATH 则从显著性预测分支中学习显著性中间特征。

总之，我们的模型在 Salicon 数据集上表现出良好的结果，尤其在 $MultiMatch$ 指标的某些组成部分上表现出色，并在 NSS 指标上取得最佳表现，体现了其准确捕捉视觉注意模式的能力。然而，$Congruency$ 指标仍有提升空间，在该指标上，其他更依赖显著性图的模型优于我们的方法。

表2展示了在MIT1003数据集上的扫描路径预测结果。我们的模型在 $Shape$ 和 $Position$ 标准上取得了最高分，这表明与其他模型相比，我们的模型在准确预测扫描路径的形状和位置方面表现出色。尽管我们的模型在 $Direction$ 和 $Length$ 特征方面的表现略低，但与最先进的方法相比仍具有竞争力。值得注意的是，我们的模型在平均分方面优于其他模型，该分数记为 $MMScore$。

在 NSS 和 $Congruency$ 分数方面，我们的模型略有下降，但仍优于对比模型。 NSS 分数反映了模型与显著区域的对齐能力，而 $Congruency$分数则体现了预测扫描路径与观测扫描路径之间的一致性。

总体而言，我们的模型在 MIT1003数据集上表现出良好的结果，在各种扫描路径特征方面均具有竞争力。它在准确捕捉扫描路径的形状和位置方面表现出色，使其成为当前先进方法中的有力竞争者。此外，该模型与显著区域和注视模式保持了良好的对齐程度，展现出其在静态图像中再现类人注意焦点的能力。

4.5. 定量组件消融

本节介绍定量组件消融研究的结果，旨在分析模型中的不同组件对扫描路径预测性能的影响。表3总结了在 MIT1003数据集上进行消融研究的结果。

4.5.1 先验图消融

我们研究了先验图和注意力偏置对我们模型性能的影响。我们将完整模型（包含先验图和注意力偏置）的结果，与移除这些组件后的模型变体进行比较。结果显示在表3中标记为“我们的模型（先验）”的行中。

模型	形状	方向	长度	位置	MM 分数	NSS	一致性
我们的模型	**0.9392**	**0.6152**	**0.9100**	**0.8537**	**0.82952**	**0.8888**	**0.2114**
我们的模型（先验）	0.9372	0.5850	0.9021	0.8519	0.8188	0.8747	0.2007
我们的模型（选择）	0.9216	0.5831	0.9087	0.8111	0.8061	0.9563	0.2326

表3：MIT1003 上扫描路径预测的消融实验结果。

与完整模型相比，移除先验图和注意力偏置后，模型在所有指标上的性能均略有下降。这表明，先验图和注意力偏置有助于模型生成准确扫描路径的能力。这些组件帮助引导模型将注意力集中到输入图像中的重要区域，从而更准确地预测类似人类的注视模式。

4.5.2 注视点选择器消融实验

为分析选择性注意机制的重要性，我们将完整模型与排除注视点选择器的模型进行性能比较。结果显示在标记为“我们的模型（选择）”的行中。没有选择性注意机制时，$MMScore$中的性能显著下降，这表明注视点选择器在提升模型捕捉类人注视模式的能力方面发挥着关键作用。

另一方面，我们注意到 NSS 和$Congruency$ 值有所增加。这是由于该组件的简单性质所致。排除该模块实际上会增加生成的注视点数量，从而提高这些指标，因为这些指标与注视点的数量及其位置直接相关。

图 4 展示了扫描路径长度的分布。图 4b 显示了原始分布，而图 4c展示了预测分布。值得注意的是，两种分布的形状极为相似且高度一致。然而，显然，与原始分布相比，预测分布中的扫描路径长度变化幅度较小。需要强调的是，我们的模型并未在该数据集上进行训练。这使得两种分布之间的高度相似更具意义。这表明，我们的模型能够很好地泛化到未见数据，并能在不同场景下有效预测扫描路径长度。

总之，这项消融研究凸显了注视点选择器组件在我们模型中的重要性。由于缺少注视点选择器会影响模型性能，因此它对于实现扫描路径预测的最佳整体性能至关重要。

指标	0.0	0.8	2.0	5.3	10.0
NSS	**1.426576**	1.175942	1.003903	0.839459	0.789592
一致性	**0.304958**	0.256472	0.224562	0.193916	0.185359
形状	**0.972415**	0.964157	0.959167	0.952888	0.951019
方向	0.537310	0.537695	0.537575	0.539900	**0.544388**
长度	**0.945042**	0.928527	0.902239	0.918545	0.905984
位置	**0.467996**	0.462994	0.459152	0.455303	0.453116
Mean	**0.7306**	0.7233	0.7145	0.7166	0.7136

表 4：不同温度下使用均匀噪声在 MIT1003 数据集上的扫描路径预测结果

指标	0.0	0.8	2.0	5.3	10.0
NSS	**0.3461**	0.2910	0.2643	0.2562	0.2512
一致性	**0.1077**	0.0982	0.0949	0.0914	0.0921
形状	**0.9761**	0.9685	0.9627	0.9542	0.9516
方向	0.5423	0.5423	0.5419	0.5440	**0.5485**
长度	**0.9525**	0.9372	0.9256	0.9086	0.9034
位置	**0.5593**	0.5509	0.5467	0.5427	0.5405
Mean	**0.7576**	0.7497	0.7442	0.7374	0.7360

表 5：基于带高斯噪声的温度设置，在 MIT1003 数据集上进行扫描路径预测的结果

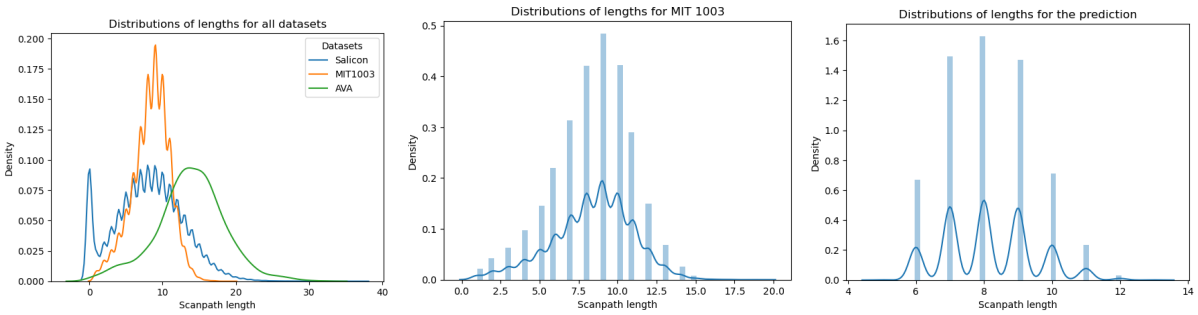
4.6. 生成网络的定量评估

表4展示了我们的扫描路径预测模型在MIT1003数据集上的结果，其中推理过程中均匀噪声的温度值各不相同。该表列出了不同温度设置下的评估分数，包括 0。0，0。8，2。0， 5。3和 10。0。

对于 NSS 指标，在温度为 0.0 时得分最高，这表明仅添加了极少量的噪声 ϵ。随着温度值升高，NSS 得分逐渐降低，说明更高水平的噪声会对模型与显著区域对齐的能力产生负面影响。

就 $Congruency$ 指标而言，也观察到了类似趋势。模型在温度为 0.0 时表现最佳；随着温度升高，$Congruency$得分降低，表明预测注视点与观测注视点之间的空间重叠和相似性有所下降。

分析该指标的各个组成部分



(a) 所有数据集的长度分布　　(b) 原始MIT1003　　(c) 预测的MIT1003

图4：MIT1003上扫描路径长度的分布。

$MultiMatch$（形状、方向、长度和位置），可以观察到，对于这些特征中的大多数，温度设为 0.0 时性能最佳。不过，对于方向和位置这两个组成部分，较高的温度值会带来略微提升的得分。

总体平均得分代表 $MultiMatch$指标的平均值，并随着温度升高而逐渐下降。这表明，在推理过程中引入的噪声越多，模型性能越会下降。

总之，表4 中的结果突出了温度和均匀噪声对模型性能的影响。较低的温度值（噪声较少）通常能更好地对齐显著区域并提高预测准确率，而较高的温度值（噪声较多）则会导致性能下降。这些发现强调了在推理过程中谨慎选择温度参数的重要性，以便在扫描路径预测中取得最佳结果。

表5 中的结果显示，与 Salicon 数据集评估（表4）中观察到的发现相似。在这两种情况下，通过温度参数化引入高斯噪声都会影响模型性能，这一点可由 NSS 和 $Congruency$ 分数体现。

对于这两个数据集，NSS和 $Congruency$ 方面的最佳性能均在温度为 0.0 时取得，这表明仅添加少量噪声能使结果与显著区域及观测注视点达到最高程度的匹配。随着温度升高（例如 0.8、2.0、5.3 和 10.0），模型性能逐渐下降，与显著性图和观测注视点的匹配程度也随之降低。

此外，$MultiMatch$指标的各个组成部分（形状、方向、长度和位置）在两个数据集中都呈现出相似趋势。较低的温度值通常能在大多数特征上带来更好的性能，但方向分量存在一些轻微差异。

此外，总体平均值分数（代表平均性能）的 $MultiMatch$指标随温度升高而持续下降，表明高斯噪声会对模型的预测准确率产生负面影响。

两个数据集中的趋势相似，进一步凸显了在推理过程中选择合适温度参数的重要性，以获得最佳性能，并提升模型在该任务上的预测泛化能力。

4.7. 绘画作品定量评估

模型	NSS	一致性
我们的模型（无领域自适应）	1.3620	0.4024
我们的模型（有领域自适应）	**1.5093**	**0.4244**

表6：领域自适应前后在 Le Meur 绘画作品上的扫描路径预测结果。

表6展示了在确定性模型中引入领域自适应后，我们的模型在 Le Meur 绘画数据集上的评估结果。由于该数据集缺少扫描路径，我们使用混合指标 NSS 和 $Congruency$ 进行评估。

NSS 指标的得分在引入领域自适应后显著提高，这表明模型预测的注视点与绘画图像中的显著区域更加吻合。这一改进体现了领域自适应在提升模型捕捉绘画领域视觉注意模式的能力方面的有效性。

此外，我们还观察到$Congruency$ 结果有所改善。$Congruency$ 指标衡量预测扫描路径与观测到的注视点之间的空间重叠；这一改善表明这两个不同的数据集更加趋于一致。这进一步证明了领域自适应在弥合自然场景图像与绘画

数据集	形状	方向	长度	位置	MM 分数	NSS	一致性
不使用领域自适应	0.9548	**0.6198**	0.9363	0.7872	0.8245	0.64761	0.1212
使用领域自适应	**0.9697**	0.6173	**0.9384**	**0.7992**	**0.8312**	**0.6534**	0.1854

表 7：我们模型在 AVAtt 绘画数据集上进行领域自适应前后的扫描路径预测结果。

图像之间差距方面的有效性，使模型能够更有效地将其表现泛化到绘画领域。

总之，表6 中的结果凸显了领域自适应对我们模型在 Le Meur 绘画数据集上评估时模型性能的积极影响。增加的 NSS 分数表明模型与显著区域的吻合度更高，而改进的 $Congruency$ 结果则表明，预测扫描路径与观测到的注视点更加吻合。这些发现凸显了领域自适应对于增强模型向新颖、多样的图像域（例如绘画作品）泛化能力的重要性。

表7 展示了在对确定性模型应用领域自适应方法前后，在 AVAtt 数据集上测试模型所获得的结果 [60] 。

表7比较了我们的模型在 AVAtt 数据集上进行领域自适应（DA）前后的扫描路径预测结果。起初，模型在形状、方向、长度、位置、MM 分数、NSS 和一致性等多项指标上的表现中等。然而，应用领域自适应后，模型性能有了显著提升。形状、长度、位置、MM 分数和 NSS 指标均大幅提高，一致性也有了显著改善。这些结果表明，领域自适应能够有效提升模型预测扫描路径和显著性值的准确性，使其在现实应用中更加可靠且适用。

4.8. 定性结果

在图5中，我们对模型生成的结果进行了定性可视化。图中展示了叠加在图像上的预测扫描路径，以及真实显著性图。每一列对应一个样本，展示了如前文所述采用不同温度时获得的结果。

所展示的结果表明，我们的模型能够高效且有效地根据输入图像预测合理的扫描路径。值得注意的是，随着温度参数升高，中心偏置的影响会减弱。因此，在较高温度下预测的扫描路径会分布在刺激的更大区域内。另一方面，在较低温度下预测的扫描路径，其注视点往往更集中于中心附近。这一观察凸显了温度对预测扫描路径的分布范围和密度的影响，为理解模型在不同条件下的行为提供了宝贵的见解。总体而言，这些定性结果证实了模型能够根据视觉刺激生成有意义且准确的扫描路径，使其成为扫描路径预测任务的一种有前景的方法。

在图**??**中，我们以直观且定性的方式展示了在新的绘画作品域上生成扫描路径的领域自适应过程。第一行展示领域自适应前生成的扫描路径，第二行则展示应用领域自适应后得到的扫描路径。

领域自适应前，扫描路径的扫视长度较短，注视点集中在有限区域内，而该区域未必对应显著区域。这种行为可能是绘画风格的影响所致，因为绘画风格会显著影响扫描路径的生成。

然而，经过领域自适应后，我们观察到扫描路径的质量显著提升。自适应后的扫描路径覆盖了更广阔的区域，并且与最后一行所示的底层显著性图高度相关。这一改进验证了领域自适应方法的有效性：它增强了模型生成扫描路径的能力，使其能够更好地契合新绘画作品域中的显著性模式。

总体而言，视觉对比展现了领域自适应的优势：即使面对新的、多样化的绘画风格，也能使模型生成更准确且更符合语境的扫描路径。

在图6 和图7中，我们分别展示了从 Le Meur 和 AVAtt 数据集获得的结果样本。这些可视化结果与先前 MIT1003 的结果相似，并展示了我们的模型在各种艺术运动和风格上的表现。

从这些图像中可以明显看出，我们的模型能够为每幅绘画作品生成多条合理的扫描路径。这些扫描路径有效地捕捉了艺术作品的显著区域，预测的注视点与绘画作品中的重要元素高度吻合。

此外，我们的模型展现出很强的通用性，在各种艺术运动和风格的绘画作品中均表现出色，涵盖文艺复兴和巴洛克艺术运动，以及古代中国水墨画。这凸显了我们模型的稳健性与适应性，使其能够处理多样化的艺术表现形式，并在不同视觉领域中生成准确的扫描路径。

数据集	形状	方向	长度	位置	MM 分数	NSS	一致性
AVAtt 数据集	0.9883	0.5046	0.9767	0.3718	0.7104	0.6323	0.1131
Le Meur数据集	-	-	-	-	-	1.2145	0.3523

表8：我们的随机模型在绘画数据集上添加高斯噪声后的扫描路径预测结果。

总而言之，Le Meur 和 AVAtt 数据集上的定性结果进一步验证了我们的模型在预测各种绘画作品和艺术风格中合理且由显著性驱动的扫描路径方面的有效性和泛化能力。

5. 结论

本文提出了一种新颖的深度学习架构，用于预测二维图像中的扫描路径，并特别关注如何将该方法应用于绘画图像——它们是文化遗产的重要组成部分。我们的模型由特征提取器、可学习的任务导向先验图，以及注视点选择模块组成，后者用于在扫描路径长度预测中引入随机性。我们还引入了一种温度调制机制，使预测的扫描路径具有多样性，以体现扫描路径的主观性。

我们将所提出的模型与最先进的方法进行了比较，发现该模型表现更优，在扫描路径预测方面取得了最佳结果。此外，我们证明了无监督领域自适应能够将模型从自然场景域适配到绘画作品域，从而为绘画作品生成高度相关的扫描路径预测。

然而，我们的模型仍有局限，目前还无法预测注视点的持续时间。未来，我们计划通过引入时间预测模块来解决这一局限。

总体而言，我们模型的领域自适应能力为文化遗产理解领域中的进一步应用开辟了道路，例如分析纪念碑、建筑、雕像和人工三维图像，以支持虚拟博物馆参观及其他与遗产相关的任务。

致谢

由TIC-ART项目及区域基金（中央-卢瓦尔河谷大区）资助

参考文献

[1] 使用脑电图并同步进行远程眼动追踪，研究内隐注意与外显注意之间的神经差异。10 592，2016年2: 。 [2] Ilyass Abouelaziz、Aladine Chetouani、Mohammed El Hassouni、Lj Latecki 和 Hocine Cherifi。三维视觉显著性以及用于盲网格质量评估的卷积神经网络。Neural Computing and Applications，2019年。2[3] Ilyass Abouelaziz、Aladine Chetouani、Mohammed El Hassouni、Longin Jan Latecki 和 Hocine Cherifi。通过卷积神经网络集成与紧凑多线性池化实现无参考网格视觉质量评价。《模式识别》，100:107174，2020年。2[4] Marc Assens、Xavier Giro-i Nieto、Kevin McGuinness 和 Noel E O' Connor。Pathgan：利用生成对抗网络进行视觉扫描路径预测。载于《欧洲计算机视觉会议论文集》（ECCV），第 0 0 页，2018年 4，12 – 。[5] Marc Assens Reina、Xavier Giro-i Nieto、Kevin McGuin- ness 和 Noel E O'Connor。Saltinet：利用显著性体对360度图像进行扫描路径预测。载于 IEEE 国际计算机视觉会议研讨会论文集，第 2331 2338, 2017 4 – 页。Wentao [6] Bao 和 Zhenzhong Chen。基于深度卷积扫视模型的人类扫描路径预测。《神经计算》，2020年。4，12[7] Neil Bruce 和 John Tsotsos。基于信息最大化的显著性。载于《神经信息处理系统进展》，第155–162页，2006年。4[8] Alessandro Bruno、Francesco Gugliuzza、Edoardo Ardiz- zone、Calogero Carlo Giunta 和 Roberto Pirrone。通过显著区域分割增强图像内容，以帮助色觉缺陷者。《i-Perception》，10(3):2041669519841073，2019年。2[9] Alessandro Bruno、Francesco Gugliuzza、Roberto Pirrone 和 Edoardo Ardizzone。一种基于多尺度颜色和关键点密度的视觉显著性检测方法。IEEE Access，8:121330–121343，2020年。4[10] Alessandro Bruno、Marouane Tliba、Mohamed Amine Kerk- ouri、Aladine Chetouani、Carlo Calogero Giunta 和 Arzu Ç¨oltekin。通过基于网络摄像头的眼动追踪检测色觉缺陷：案例研究。载于2023年眼动追踪研究与应用研讨会论文集，ETRA '23，美国纽约，2023年1, 2美国计算机协会。 [11] Aladine Chetouani。一种基于卷积神经网络选择相关图像块的盲图像质量度量。载于 EUSIPCO，第 1452 1456, 2018 2 – 页。 [12] Aladine Chetouani。用于盲图像质量评价的卷积神经网络与显著性选择。载于2018年IEEE第25届国际图像处理会议（ICIP），第 2835 2839 2018 2 – 页。IEEE，。 [13] Aladine Chetouani 和 Leida Li。关于使用扫描路径预测器和卷积神经网络进行盲图像质量评估。《信号处理：图像通信》，89:115963，2020年。2

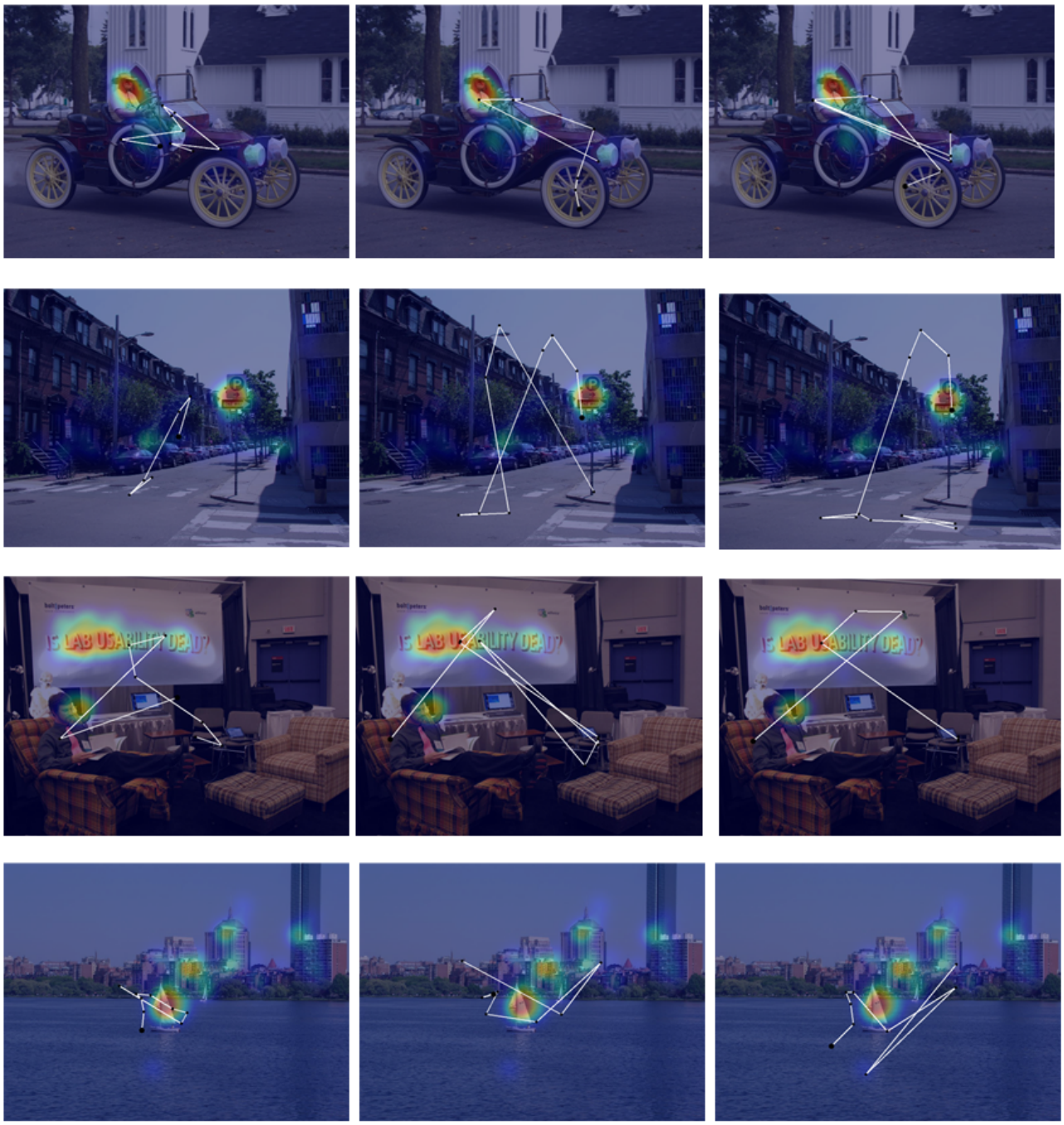


图5：MIT1003 上的定性预测。

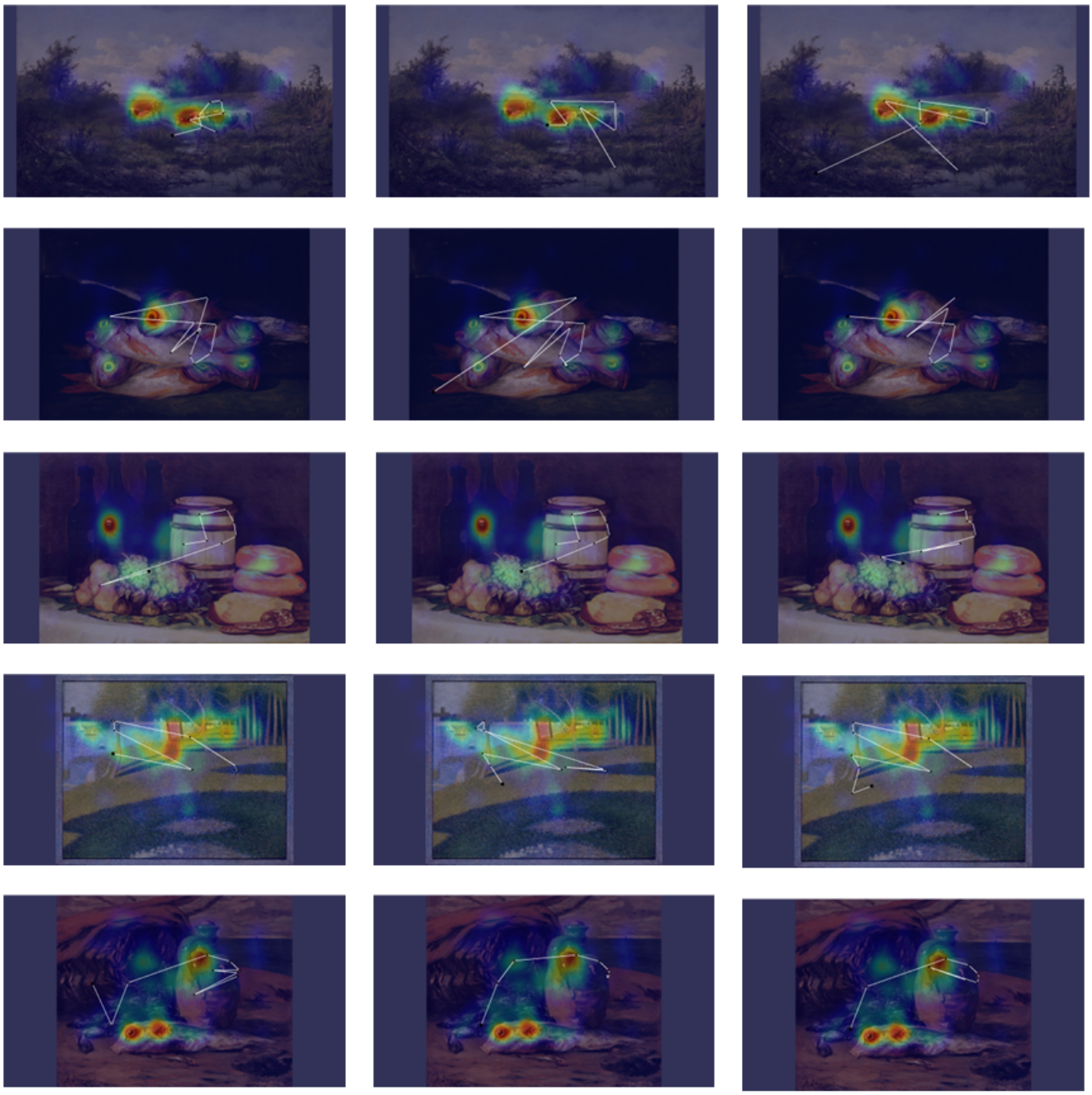


图6：Le Meur数据集上的定性预测。

[14] Marcella Cornia、Lorenzo Baraldi、Giuseppe Serra 和 Rita Cucchiara。用于显著性预测的深度多层网络。载于 *2016*年 第*23*届国际模式识别会议 *(ICPR)*，第 3488–3493 页。IEEE，2016年。 4

[15] Marcella Cornia、Lorenzo Baraldi、Giuseppe Serra 和 Rita Cucchiara。用于新型多媒体服务中图像描述的视觉显著性。多媒体服务。*2017*年 *IEEE*多媒体与博览会国际会议研讨会（*ICMEW*），第 309–314 页。IEEE，2017年。2

[16] Marcella Cornia、Lorenzo Baraldi、Giuseppe Serra 和 Rita Cucchiara。通过基于LSTM的显著性注意模型预测人眼注视点。*IEEE* 图像处理汇刊

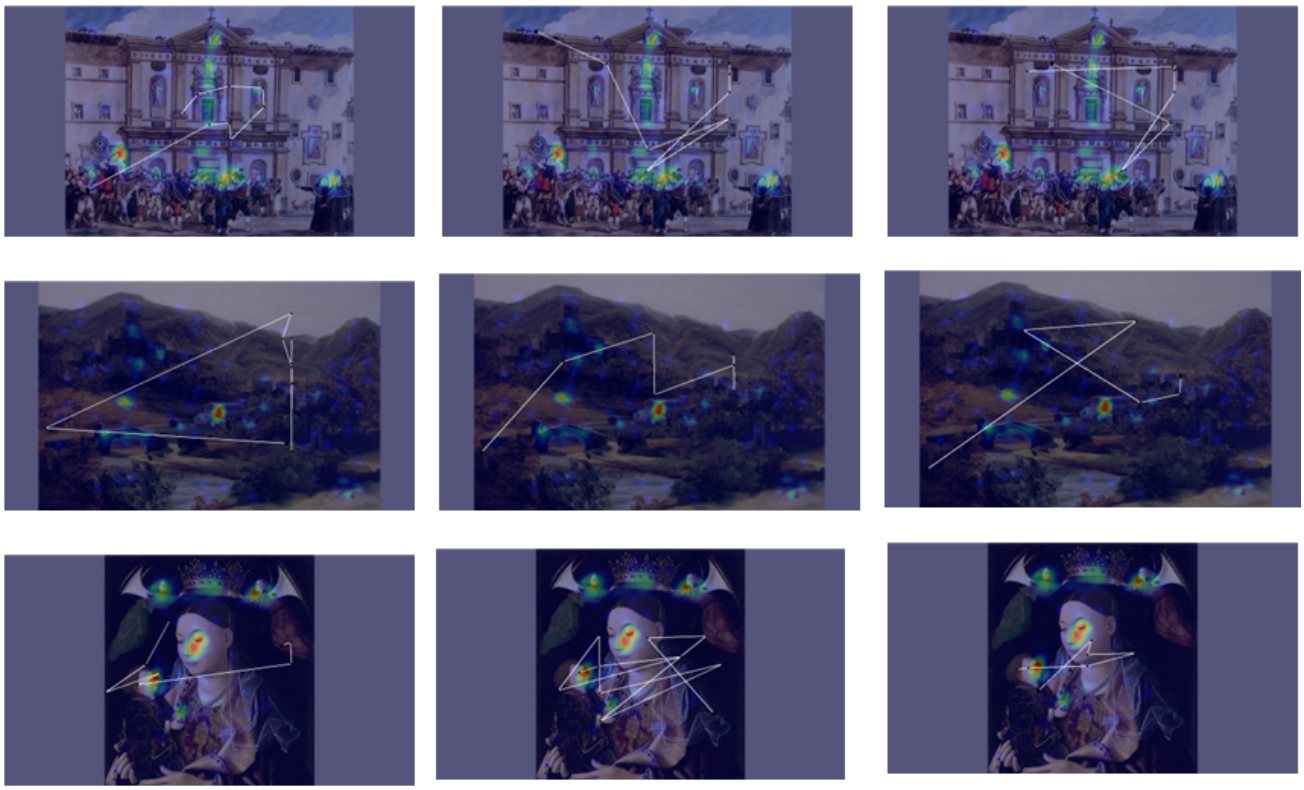


图7：AVAtt 数据集上的定性预测。

cessing，27(10):5142–5154，2018年。4，6

[17] Ryan Anthony Jalova de Belen、Tomasz Bednarz 和 Ar- cot Sowmya。Scanpathnet：一种用于扫描路径预测的循环混合密度网络。载于 *2022 IEEE/CVF* 计算机视觉与模式识别会议研讨会（*CVPRW*），第 5006–5016 页，2022年。4[18] Richard Dewhurst、Marcus Nystr¨om、Halszka Jarodzka、Tom Foulsham、Roger Johansson 和 Kenneth Holmqvist。观其所见：使用 MultiMatch（一种基于向量的方法）进行多维扫描路径比较。《行为研究方法》，44(4):1079–1100，2012年。10，11[19] Richard Droste、Jianbo Jiao 和 J Alison Noble。统一的图像与视频显著性建模。*arXiv* 预印本 *arXiv:2003.05477*，2020年。4[20] M Hamidi 等。基于网格显著性和小波变换的盲鲁棒三维网格水印，用于版权保护。《*Information*》，10(2)，2019年。2[21] Kunihiko Fukushima。Neocognitron：一种自组织神经网络模型，用于实现不受位置偏移影响的模式识别机制。《生物控制论》，36(4):193–202，1980 年。5[22] Yaroslav Ganin、Evgeniya Ustinova、Hana Ajakan、Pas- cal Germain、Hugo Larochelle、Franc.ois Laviolette、Mario March 和 Victor Lempitsky。域对抗训练

的神经网络。《机器学习研究杂志》，17(59):1–35，2016年。4，7[23] Joseph H. Goldberg 和 Jonathan I. Helfman。视觉扫描路径表示。在*2010*年眼动追踪研究与应用研讨会论文集，ETRA '10，第203–210页，纽约，纽约州，美国，2010年。美国计算机协会。2[24] Chenlei Guo、Qi Ma 和 Liming Zhang。使用四元数傅里叶变换的相位谱进行时空显著性检测。在*2008*年*IEEE*计算机视觉与模式识别会议，第1–8页。IEEE，2008年。4[25] Jonathan Harel、Christof Koch 和 Pietro Perona。基于图的视觉显著性。在《神经信息处理系统进展》，第545–552页，2007年。4[26] Rudi Hartmann。遗产与经济。载于 Audrey Kobayashi 编，《国际人文地理百科全书》（第二版），第369–372页。爱思唯尔，牛津，第二版，2020年。1[27] Huiting Hong、Xin Li、Yuangang Pan 和 Ivor W. Tsang。域对抗网络对齐。《*IEEE*知识与数据工程汇刊》，34 3224，2022年 7 3211*3*( ): – 。[28] Laurent Itti 和 Christof Koch。视觉注意的计算建模。《自然综述：神经科学》，2 203，2001年 3 194*3*，*4*( ): – 。

[29] Ming Jiang、Shengsheng Huang、Juanyong Duan 和 Qi Zhao。Salicon：上下文显著性。载于 CVPR，第 1072–1080页。IEEE计算机学会，2015年。4，7，8，9，11[30] Jost B Jonas、Ulrike Schneider 和 Gottfried OH Naumann。人类视网膜感光细胞的数量与密度。《Graefe眼科临床与实验档案》，230(6):505–510，1992年。5[31] Tilke Judd、Fr´edo Durand 和 Antonio Torralba。用于预测人类注视点的显著性计算模型基准。载于 MIT技术报告，2012年。9[32] Tilke Judd、Krista Ehinger、Fr´edo Durand 和 Antonio Torralba。学习预测人类注视的位置。载于 IEEE国际计算机视觉会议（ICCV），2009年。4，8，11[33] Fumi Katsuki 和 Christos Constantinidis。自下而上与自上而下注意：不同的过程与重叠的神经系统。《神经科学家》，20(5):509–521，2014年。PMID：24362813。2[34]Mohamed Amine Kerkouri、Marouane Tliba、Aladine Chetouani 和 Alessandro Bruno。用于绘画作品扫描路径预测的领域自适应深度学习解决方案。载于第19届国际基于内容的多媒体索引会议论文集，CBMI ’22，第57–63页，纽约，纽约州，美国，2022年。美国计算机协会。4[35]Mohamed Amine Kerkouri、Marouane Tliba、Aladine Chetouani 和 Alessandro Bruno。用于视觉扫描路径预测的观察者间一致深度对抗训练，2023年。4[36] Mohamed Amine Kerkouri、Marouane Tliba、Aladine Chetouani 和 Alessandro Bruno。AVAtt：适用于多样绘画风格的艺术视觉注意数据集。载于2024年眼动追踪研究与应用研讨会论文集，ETRA ’24，纽约，纽约州，美国，2024年。美国计算机协会。8，9[37] Mohamed A. Kerkouri、Marouane Tliba、Aladine Chetouani 和 Rachid Harba。SALYPATH：用于视觉注意预测的基于深度学习的架构。载于2021年IEEE国际图像处理会议（ICIP），第1464–1468页，2021年。4，6，12[38] Christof Koch 和 Shimon Ullman。选择性视觉注意的转变：探寻其背后的神经回路。载于《智能问题》，第115–141页。Springer，1987年。2，3[39] Matthias K¨ummerer、Lucas Theis 和 Matthias Bethge。DeepGaze I：利用基于ImageNet训练的特征图提升显著性预测。arXiv预印本 arXiv:1411.1045，2014年。4[40]Matthias K¨ummerer、Thomas SA Wallis 和 Matthias Bethge。DeepGaze II：从经过目标识别训练的深度特征中读取注视点。arXiv预印本 arXiv:1610.01563，2016年。4[41] Matthias K¨ummerer、Thomas Wallis 和 Matthias Bethge。DeepGaze III：利用深度学习探究场景内容与扫描路径历史在注视点选择中的相互作用。2019年1月。4[42] Olivier Le Meur、Thierry Baccino 和 Aline Roumy。观察者间视觉一致性（IOVC）的预测及应用。

化到图像排序。在第19届美国计算机协会国际多媒体会议论文集中，页373–382，2011年。10[43] Olivier Le Meur、Tugdual Le Pen 和 R´emi Cozot。我们能否准确预测人们观看绘画作品时注视的位置？PLOS ONE，15：1–20，10 2020年。8，9[44] Olivier Le Meur 和 Zhi Liu。自由观看条件下的眼球运动扫视模型。Vision research，116：152–164，2015年。4，12[45] Y. Lecun、L. Bottou、Y. Bengio 和 P. Haffner。应用于文档识别的基于梯度的学习。IEEE论文集，86(11)：2278–2324，1998年。5[46] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：情境中的常见物体。载于欧洲计算机视觉会议论文集，页740–755。Springer，2014年。8[47] Jianxun Lou、Hanhe Lin、David Marshall、Dietmar Saupe 和 Hantao Liu。TransalNet：迈向符合感知的视觉显著性预测。Neurocomputing，494 467，2022 455 4：–。[48] Nobuyuki Otsu。一种基于灰度级直方图的阈值选择方法。IEEE系统、人类与控制论汇刊，9(1)：62–66，1979年。11[49] Junting Pan、Cristian Canton Ferrer、Kevin McGuinness、Noel E O’Connor、Jordi Torres、Elisa Sayrol 和 Xavier Giro-i Nieto。SalGAN：利用生成对抗网络进行视觉显著性预测。arXiv预印本 arXiv:1701 01081，2017 4。。[50] Junting Pan、Elisa Sayrol、Xavier Giro-i Nieto、Kevin McGuinness 和 Noel E O’Connor。用于显著性预测的浅层与深层卷积网络。载于IEEE计算机视觉与模式识别会议论文集，页598–606，2016年。4[51] Yash Patel、Srikar Appalaraju 和 R Manmatha。显著性驱动的感知图像压缩。载于IEEE/CVF冬季计算机视觉应用会议论文集，页227–236，2021年。2[52] Robert Peters 和 Laurent Itti。傅里叶相位信息在预测显著性中的作用。Journal of Vision，8(6)：879–879，2008年。4[53] Robert J Peters、Asha Iyer、Laurent Itti 和 Christof Koch。自然图像中自下而上的注视分配的组成部分。Vision research，45(18)：2397–2416，2005年。10[54] Marcus E Raichle。大脑的暗能量。Scientific American，302(3)：44–49，2010年。1[55] Mark Sandler、Andrew Howard、Menglong Zhu、Andrey Zhmoginov 和 Liang-Chieh Chen。MobileNetV2：倒置残差与线性瓶颈。载于IEEE计算机视觉与模式识别会议论文集，页4510–4520，2018年。5[56] Leo Schwinn、Doina Precup、Bjoern Eskofier 和 Dario Zanca。利用神经视觉注意力模拟人类注视。载于NeuRIPS 2022 Gaze Meets ML研讨会，2022年。4[57] Xuan Shao、Ye Luo、Dandan Zhu、Shuqin Li、Laurent Itti 和 Jianwei Lu。基于高层特征和记忆偏差的扫描路径预测。载于国际会议

神经信息处理，页3–13。Springer，2017年。4[58] Xiangjie Sui、Yuming Fang、Hanwei Zhu、Shiqi Wang 和 Zhou Wang。ScanDMM：面向360度图像扫描路径预测的深度马尔可夫模型。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，页6989–6999，2023年6月。5[59] Marouane Tliba 等。Satsal：一种基于多层级自注意力的视觉显著性预测架构。IEEE Access，页1–1，2022年。2，4[60]Marouane Tliba、Mohamed Amine Kerkouri、Aladine Chetouani 和 Alessandro Bruno。面向绘画图像的自监督扫描路径预测框架。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）研讨会论文集，页1539–1548，2022年6月。2，4，8，9，15[61]Marouane Tliba、Mohamed Amine Kerkouri、Aladine Chetouani、Alessandro Bruno、Mohammed El Hassouni 和 Arzu C ¸ ¨oltekin。通过基于眼动追踪的策略对掩码自编码器涌现属性进行感知评估。载于2024年眼动追踪研究与应用研讨会论文集，ETRA '24，美国纽约，2024年。美国计算机协会。2[62] Marouane Tliba、Aymen Sekhri、Mohamed Amine Kerkouri 和 Aladine Chetouani。通过领域自适应进行基于深度学习的医学图像质量评估。载于2022年IEEE国际图像处理会议（ICIP），页3692–3696，2022年。4[63] Anne M Treisman 和 Garry Gelade。注意的特征整合理论。认知心理学，12 136，1980 1 97 1（）：–。[64] H. von Helmholtz 和 J.P.C. Southall。《亥姆霍兹生理光学论著》。眼科学经典文库第3卷。美国光学学会，1925年。1[65] Haoxiang Wang、Zhihui Li、Yang Li、Brij B Gupta 和 Chang Choi。视觉显著性引导的复杂图像检索。模式识别快报，130：64–72，2020年。2[66] Yao Wang、Mihai Bˆ ace 和 Andreas Bulling。信息可视化中的扫描路径预测，2023年。4[67] Yin Yue 和 Zhou Jun。基于显著性检测和TV模型的岩画修复方法。载于2021年国际计算机工程与应用会议（ICCEA），页122–125，2021年。2[68] Dario Zanca、Stefano Melacci 和 Marco Gori。注意焦点的引力定律。IEEE模式分析与机器智能汇刊，2019年。4，12[69] Xilin Zhang、Li Zhaoping、Tiangang Zhou 和 Fang Fang。V1中的神经活动产生自下而上的显著性图。Neuron，73(1)：183–192，2012年。2