IC9600: 用于图像复杂度自动评估的基线数据集

冯停磊，翟英杰，杨巨峰，梁杰，范登平，张敬，邵岭，陶大程

摘要—图像复杂度 (image complexity, IC) 是人类理解图像时的一种重要视觉感知。精确地评估 IC 具有挑战性，而且长期以来一直被忽视。一方面 IC 的评估因依赖于人类的感知而相对主观，另一方面 IC 依赖于语义，而现实世界的图像是多样化的。为了促进在深度学习时代的 IC 评估研究，本文构建了已知的第一个包含 9600 张良好标注图像的 IC 评估数据集。这些图像涉及到多种不同领域，如：抽象画、绘画、真实世界场景等等，每一张图像都由 17 位人类标注者精心标注。在这个高质量数据集的支持下，本文进一步提供了一个基础模型来预测 IC 分数，并以弱监督的方式估计复杂度热度图。该模型在 IC 评估中表现出良好的性能，它的预测结果与人类感知的相关性（皮尔逊相关系数）达到了 0.949。最后，本文通过实验验证了 IC 可以提供有效的辅助信息，并广泛提升多种计算机视觉任务的性能。数据集和代码可见于：https://github.com/tinglyfeng/IC9600

Index Terms—图像复杂度评估，图像属性，大规模精细标注数据集

1 引言

图像复杂度 (image complexity, IC) 定义为一幅图像中蕴含的错综复杂性 [2]。客观而言，可以认为 IC 是图像中的细节和内容变化的数量 [3]。而从主观上来说，它是人类观察者从全局抽象和局部细节等角度理解或描述一幅图像的困难程度 [3], [4]。如图 1 所示，简单的素描和开阔的天空具有较低的 IC，而建筑的纹理和拥挤的人群则包含相对较高的 IC，一幅图像的整体 IC 受到这种具有不同 IC 强度的局部区域及组合的影响。IC 是心理学中的一种重要感知，它能强烈影响视觉美学 [5], [6] 和观看者的情感反应 [7]，同时也是计算机视觉中一种常见的重要属性。IC 的自动评估已被证明有助于多种应用，如图像分割 [8]、图像隐写术 [9]、网页设计 [7]、文字检测 [10]、图像增强 [11] 等等。因此，为了模仿人类的复杂度感知以及促进这些相关任务的发展，精确的 IC 评估颇为重要。

为了实现这一目标，前人工作使用一些启发性指标来评估 IC，比如图像熵 [12]、原始图像与其压缩图（JPEG, GIF 等）之间的大小比例 [13]、整个图像中边缘像素的密度 [14]、图像中独特的 RGB 颜色计数 [15] 等等。除此以外，机器学习方法如支持向量机、随机森林、反向传播神经网络等 [15], [16]

图 1. 本文的 IC9600 数据集中不同类别的图像样本展示，如抽象画、真实世界场景、建筑等。'S1'-'S5' 表示由 17 位标注者标注的复杂度分数（1-5 分）的分布。这些图像按平均复杂度分数（每个图像的顶部，归一化为 $[0, 1]$）进行排序。

也通过利用各种基本图像特征来预测 IC。上述算法主要基于小规模数据集，并侧重于使用手工提取的特征，这些因素阻碍了它们在实际应用中预测图像复杂度的泛化能力。

图像复杂度评估的挑战性主要在于以下几点: (1) 现实世界的图像在几乎无限多的模式和场景中变化，因此很难根据手工提取的特征组合来鲁棒地表示其复杂度。(2) IC 是一种依靠人类感知的高层次（主观）概念，与图像底层特征有很大差距。近年来，深度卷积神经网络（convolutional neural networks, CNN）表现出强大的表示和泛化能力，并以数据驱动的方式对人类的主观感知（如图像美学 [23]，图像质量 [24] 等等）进行精确建模。然而，现有的 IC 数据集都是小规模的，多样性有限，因为构建 IC 数据集是一项艰巨而耗时的工作（每张图片都需要由足够多的标注者进行标注以减少模糊性）。如表 1 所示，这些数据集要么规模小，要么尚未公开，要么

- 前两位作者对本文贡献相同。
- 冯停磊，翟英杰，杨巨峰，范登平：南开大学计算机学院，中国。（邮件：*tinglyfeng@163.com, zhaiyingjier@163.com, yangjufeng@nankai.edu.cn, dengpingfan@mail.nankai.edu.cn*）
- 梁杰：香港理工大学，中国香港。（邮件：*liang27jie@163.com*）
- 张敬：悉尼大学计算机学院，澳大利亚。（邮件：*jing.zhang1@sydney.edu.au*）
- 邵岭：特斯联集团，中国。（邮件：*ling.shao@ieee.org*）
- 陶大程：京东探索研究院，中国，以及悉尼大学计算机学院，澳大利亚。（邮件：*dacheng.tao@gmail.com*）
- 本文通讯作者为：杨巨峰。

- 本文为 *TPAMI2023* 论文 *[1]* 的中文译版，由冯停磊翻译，范登平、杨巨峰校稿。

表 1
目前用于图像复杂度评估的数据集概述。第 7 行和第 8 行的缩写分别表示:Abs (抽象), Adv (广告), Arc (建筑), Art (艺术), Id (室内设计), Obj (物体), Pai (绘画), Per (人群), Sce (场景), Sup (至上主义绘画), Tra (交通), Vi (信息图表) 。

#	数据集	年份	大小	图像类型	标注方式	是否公开
1	Oliva 等人 [17]	2004	100	室内场景	分值 (1-8)	否
2	Corel 1000A [18]	2013	1,000	物体	三个类别	否
3	Miniukovich 等人 [19]	2014	140	网页	分值 (1-5)	否
4	Corchs 等人 [20]	2016	220	真实世界场景 (98), 真实材质 (122)	分值 (0-100)	否
5	Fan 等人 [21]	2017	40	中国水墨画	分值 (1-7)	否
6	Guo 等人 [15]	2018	500	绘画图像	分值 (1-7)	否
7	SAVOIAS [22]	2020	1,420	Adv, Art, Id, Obj, Sce, Sup, Vi	逐对比较	是
8	**IC9600 (本文)**	2023	**9,600**	Abs, Adv, Arc, Obj, Pai, Per, Sce, Tra	分值 (1-5)	是

只限于特定的主题，所以很难为有效的数据驱动和基于深度学习的图像复杂度分析方法提供动力。

为了促进深度学习时代的 IC 研究，本文构建了一个包含 9600 张图像的大规模数据集，并称之为 IC9600。每张图片都由 17 个训练有素的标注者标注，这些标注者是通过精心设计的复杂度感知测试筛选出来的。如图 1 所示，本文提出的数据集包含多样的语义类别，即抽象、广告、建筑、物体、绘画、人群、场景和交通。目前与 *IC9600* 最相关的数据集是 SAVOIAS [22]，不同点在于本文的数据集具有更大的规模、更多样化和面向应用的类别，以及在大量样本情景下更实用的标注方案。通过多样化的主题，本文的目标是支持深度和鲁棒的模型训练，并提供全面的辅助表征来促进广泛的相关任务的发展。

基于该数据集本文提出了一个基础模型 ICNet，以提取 IC 表征并促进其他应用的发展。ICNet 设计有两个分支，即细节分支和全局分支。细节分支利用浅层卷积网络从高分辨率图像中捕捉局部表征。全局分支通过更深的网络从较小尺寸的图像中挖掘出整体的背景信息。接下来，本文将来自两个分支的特征信息合并，并送到后续的两个预测头中。其中一个预测头回归 IC 分数，用来代表图像的整体复杂度，而另一个预测头输出 IC 热度图，用来描述图像的局部复杂强度。实验结果证明了本文所提方法的有效性。

为了进一步证明图像复杂度在计算机视觉中的重要意义，本文探索了先验复杂度信息与特定视觉任务（比如图像美学评估 [5]，人群计数 [25]，显著性物体检测 [26] 等等）之间的关系。在多个数据集上的实验表明，IC 可以提供有效的辅助信息，并且通过适当的方式利用 IC 信息能够提升下游任务的性能。

本文的贡献可以概括为以下几点：

- 构建了目前最大的有良好标注的 IC 数据集，它解决了对大规模 IC 评估数据集的迫切需求。本文提出的数据集包含 9600 张图像，涵盖 8 个语义类别。每个样本都是由多人进行标注的，最大限度避免了主观偏见。
- 设计了一个基线模型来预测图像的复杂度分数，并以弱监督的方式估计图像的复杂度热度图。该模型包含两个独立分支，分别提取细节和全局特征。此外，本文还提出了一个空间分布注意力模块，以进一步提升模型性能。
- 本文将提出的复杂度预测模型应用于多种计算机视觉任务，初步探索了深度学习时代 IC 的使用方式。实验结果表明，IC 作为一种重要的图像属性可以用来提高下游视觉任务的性能。

2 相关工作

2.1 图像复杂度评估

至今已经有很多研究人员从心理学角度研究了影响人类对 IC 感知的因素 [3], [27], [28]。Oliva 等人 [17] 将 IC 的表现特征描述为物体的数量、聚集、开放、对称、组织和颜色的多样性。Forsythe [2] 认为，熟悉程度是影响 IC 感知的重要因素。例如，观察者倾向于对熟悉的形状给出比实际更低的复杂度评价。Purchase 等人 [29] 进行了一项实验研究，调查 IC 是否可以被量化以及是否可以与参与者的复杂度观点相匹配。研究结果表明，定义一个明确的指标来充分衡量人类对 IC 的感知是具有挑战性的。

研究者已经开发了一些算法用于自动评估图像复杂度 [12], [30]，其中一些尝试用熵来量化 IC。Stamps [27] 研究了 IC 和熵的刺激之间的关系，发现它们的关联性很强，而且是线性的。基于图像块和视觉信息之间的关系，Rosenholtz 等人 [31] 利用特征拥塞和子带熵来测量 IC。Machado 等人 [13] 认为简单的图像往往有更多的冗余信息，而复杂图像中的像素值则更不容易被预测。换句话说，简单图像通常可以被压缩到更小，因此引入压缩率来描述 IC 是一种可行的方案。即使熵被证明与 IC 相关，它们也只能粗略地评估一个图像携带的信息量，更精确地评估 IC 需要更具体的描述工具（如图像的手工特征）。在 [32] 中，构图、色彩和内容的分布被定义为影响人类视觉感知的主要因素。因此，他们设计了 29 种局部、全局和显著区域的特征来代表上述三个因素。除此以外，边缘密度 [13]、空间信息 [33]、视觉注意力 [30] 等也被作为计算 IC 的常用指标。除此之外，一些机器学习方法也已经通过结合手工特征来为 IC 建模。例如，Sun 等人 [5] 采用梯度提升树来回归复杂度的构成、统计和分布特征。Chen 等人 [16] 使用反向传播神经网络来建立 IC 和三个特征之间的关系，即纹

理、边缘和区域。最近，Abdelwahab 等人 [34] 使用预先训练的 CNN 提取特征，然后使用支持向量机预测复杂程度。类似地，Saraee 等人 [22] 提出用岭回归预测 IC，其中输入的图像表征是从预先训练好的 CNN 中提取的。此外，通过深度神经网络中间层的激活强度来模拟图像复杂度，该工作还开发了一种无监督的激活能（unsupervised activation energy, UAE）方法来评估 IC。

即使目前已经提出了各种方法来研究 IC，这些方法大多是基于手工特征和传统的机器学习方法。最近的一些方法采用了现成的预训练好的 CNN 来提取图像特征，但由于缺乏大规模的配对训练数据，这些 CNN 无法以端到端的方式训练，因此很难匹配 IC 的高层感知属性。另外，由于缺乏基线数据集，大多数现有工作都是在自己构建的单个数据集上进行实验并报告性能，所以实验的比较可能会有所偏差。这些数据集规模小，或有主观偏见（即每个样本都是由有限的标注者标注），所以可能得不到令人信服的结论。基于上述分析，现在迫切需要一个大规模和高质量的 IC 基线数据集来进行公平的比较，以此推动 IC 评估的发展。

2.2 图像主观属性评估

图像复杂度和许多其他图像属性如图像质量 [35]、图像美学 [36] 一样，属于因人而异的主观概念。在这里，由于和 IC 评估的相似度最高，本节主要研究图像质量评估（image quality assessment, IQA）和图像美学评估（image aesthetic assessment, IAA）任务的成功之处。

在 IQA 和 IAA 领域，为了降低主观视觉属性标签中的个人偏见，对来自多人的感知数据进行平均已经被证明是当下的一个有效方法。Ponomarenko 等人 [37] 提出了一个广泛使用的 IQA 数据集 TID2008，其中表示图像质量的平均意见分数是从 $256k$ 的个人评估中收集的。最近提出的 KADID-10k 数据集 [38] 包括更多的图像和失真类型，每张图像通过众包获得 30 个失真类别评级。对于 IAA 任务，CUHK-PQ [39] 是早期的经典数据集。每张图片由 10 名观察者标注，并赋予一个二值标签表示图像美学的等级。紧随着这项工作，Murray 等人 建立的 AVA [40] 是目前最大的 IAA 数据集，其中每张图片都由多人投票打出 1 至 10 的分数。在这些数据集和人类平均感知的监督下，研究者提出了基于深度网络的模型来精确评估图像质量或美学。例如，Kang 等人 [41] 提出了由一个卷积层、两个全连接层和一个输出节点组成的网络。这个模型在 IQA 中被证明比基于手工特征的传统方法更有效。Bianco 等人 [42] 提出了 DeepBIQ 模型，它首先预测子区域的分数，然后对其进行平均来估计图像质量。考虑到放缩、裁剪或填充等图像变换可能会破坏图像的美学，Mai 等人 [43] 引入了一个成分保留网络，它可以直接处理原始尺寸的输入图像，同时提取多尺度特征。Zhang 等人 [44] 提出了一个 gated peripheral-foveal 卷积神经网络来模仿人类的审美感知机制，这个模型可以同时编码整体信息和细粒度特征。此外，图像裁剪 [45]、群体最大差异化竞争 [46], [47]、自监督特征学习 [48] 也被证实是 IAA 和 IQA 的有效方法。根据这些 IQA 和 IAA 的成功经验，本文构建了 IC9600 基线数据集，并提出了基于 CNN 的模型来同时预测 IC 分数及复杂度图。

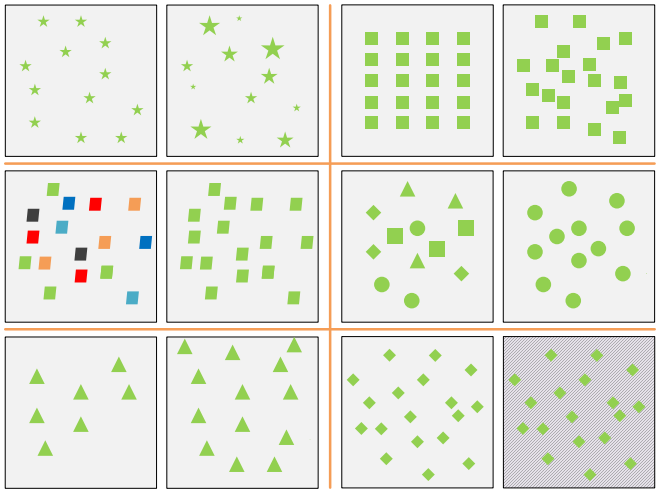

图 2. 本文的感知测试中的代表性样本。每一对图像被人工设置为不同的 IC 级别，候选人被要求在每一对中选择更复杂的图像。

2.3 图像复杂度数据集

如表 1 所示，几个小型的数据集已被用于 IC 分析。具体来说，Oliva 等人 [17] 收集了 100 张室内场景的图片，然后通过二分法（分成简单组和复杂组）对它们进行标注。Iliyasu 等人 [18] 建立了 Corel 1000A 数据集，其中的图像被标注为三个类别（即简单、普通和复杂）。为了研究用户图形界面的复杂度度量，Miniukovich 等人 [19] 收集了 140 个网页的截图，然后让 10 名研究生以 1-5 分的标准对其复杂度进行评分。之后，Corchs 等人 [20] 收集了 98 张场景图像和 122 张真实纹理图像来预测真实世界图像的复杂度感知。最近，为了确定影响绘画的 IC 感知的因素，Fan 等人 [21] 构建了一个包含 40 张中国水墨画的复杂度数据集。Guo 等人 [15] 采用 1-7 评分制收集了 500 张绘画图像。此外，Saraee 等人 [22] 创建了 SAVOIAS 数据集，其中包含超过 1000 张图像和无偏见的真实标签用于 IC 分析。

SAVOIAS 是与本文的 IC9600 最相似的数据集，但本文数据集具有以下关键优势：首先，IC9600 的规模比 SAVOIAS 大接近 7 倍。其次，IC9600 涵盖了更多主题（8 比 7），并且更贴近于现实世界的应用。最后，IC9600 的标注跨越了整个数据集而与类别无关，而 SAVOIAS 中的真实分数只在同一类别中具有可比性，这限制了其在通用图像复杂度分析中的普适应用。

可以注意到现有的与复杂度相关的数据集都是小规模的，而且大部分都未公开。因此，有必要建立一个符合真实世界特点和要求的大规模数据集，以解决通过深度学习对 IC 进行评估的迫切需求。

3 数据集

3.1 数据收集

• **图像来源** 本文的数据集包含八个类别，包括抽象、广告、建筑、物体、绘画、人物、场景和交通。为了建立这样一个多样化的数据集，本文从几个流行的数据集中收集每个类别的图像。具体来说，本文从 AVA [40] 中选择抽象和建筑图像，从 Image and Video Advertisements [49] 中选择广告图像，从 MS-COCO [50] 中选择物体图像。而绘画图片来自 JenAesthetics [36]，人物图片来自 WiderPerson [51]，场景图片来自 Places365 [52]，交通图片来自 BDD100K [53]。

• **取样策略** 为了进一步提高每个语义类别的多样性，本文从每个数据集中选择图像时尽可能多地包含子类别。例如，Places365 [52] 数据集包含 365 个场景类别。本文从每个场景类别中随机选择 4 张图像，得到总共 1460 张场景类别的图像。其他类别的采样策略与此类似。经过采样后，对于八个类别，每个类别都得到约 1500 张图像。

• **去重和过滤** 注意到一些从不同数据集采样的图像可能是相同的或非常接近的。因此，本文使用 Image Deduplicator[1] 工具去除重复的图像。之后再过滤掉那些有太多水印或质量不高的图片。经过多轮检查和选择，最后得到了总共 9600 张图像。

3.2 图像标注

• **标注指引和测试** 为了确保标注的质量，本文精心挑选、训练和测试标注者。首先，本文从与视觉相关的大学实验室中选择标注的候选人。其次，通过详细的教程对他们进行培训，包括研究的目的和 IC 的基本概念。第三，为了验证他们区分不同 IC 等级的能力，本文进行了一个感知测试。该测试包括一些用简单的几何形状（如三角形、矩形和圆形）人工生成的图像对。图像对的样本可见于图 2。每个图像对中的两幅图像根据 IC 因素，包括纹理、形状、物体排列等，被人为地设置为不同的复杂级别。根据测试结果，最终得到了 20 个合格标注者，他们在测试中表现优异（所有标注者的得分都在 90% 以上）。为了保证标注者有能力区分多个等级的 IC，本文根据 IC 的基本属性，为每个类别选择了具有 5 种复杂程度的图像（即非常简单、简单、中等、复杂和非常复杂），每个复杂度级别包括 5 张图像。本文要求标注者观察多个复杂度等级的差异，并在标注图像时将其作为参考。

• **标注环境** 每个标注者都被要求在一个安静的房间里，不受任何其他人的干扰。他们被要求在标注时集中注意力，并保持手机静音。

• **标注过程** 参考 [5], [23]，本文让每个标注者为每张图像标注 1-5 的复杂度等级，即标注范围为非常简单（1）到非常复杂（5）。每个标注者应在一天内标注 320 张图像，总共需要 30 天的时间。而且他们应该观察一张图像超过 10 秒的时间，以便对图像有一个准确的感知。如果他们觉得累了，就应该停止标注，进行适当的休息。此外，每次在标注之前，他们都需要回顾 IC 概念和多个 IC 级别的区别。这里注意，应用于 SAVOIAS 数据集 [22] 的配对比较的标注方法并不适用于本文的数据集，因为：(1) 与 SAVOIAS（每个类别分别标注约 200 张图像）不同，本文的数据集要大得多（9600 张图像）。因此，配对比较的工作量将呈指数级增长，这超出了标注者的能力范围。(2) 多等级标注法在以前许多与图像质量评估 [35] 和图像美学评估 [23] 等图像主观属性评估有关的工作中被验证是可靠的。此外，多等级标注法的主观偏差可以通过平均的策略得到抑制。

1. https://github.com/idealo/imagededup

表 2
不同组合设置的平均 PCC。等式 $M\times N+K$（第 1 行）表示本文将 17 个标注分成 $N+1$ 组，N 组中的每组包含 M 个标注，单独的 1 组包含 K 个标注。之后，每组中的多个标注被平均为 1 个标注。对于这 $N+1$ 组，计算所有 C_{N+1}^2 个组合对的 PCC。第 2 行是 C_{N+1}^2 个 PCC 的平均值。

分组	1×17	$2\times7+3$	$3\times4+5$	$4\times3+5$	$5\times2+7$	$8\times1+9$
PCC	0.54	0.68	0.77	0.82	0.86	0.94

• **离群标注** 参考以前的工作 [54]，可以认为与其他标注者有非常低的一致性的标注者是离群的。具体来说，对于每个标注者，本文计算他与其他标注者的平均 PCC。最后，本文将平均 PCC 低于 0.4 的标注者作为离群值移除，最终得到 17 个标注来计算复杂度得分。

• **移除主观性** 本文利用广泛使用的平均策略来减少标注者的主观性。特别地，经过标注之后，本文定义第 j 个样本的标签集合为 $\{y_j^1, y_j^2, y_j^3, \cdots y_j^m\}$，这里 $y_j^i \in \{1,2,3,4,5\}$，同时 m 是为第 j 个图像分配的所有标签数量。最终的复杂度分数 l^j 是通过对 m 个标签进行平均获得的，然后再归一化为 $[0,1]$，即 $l^j == \sum_{i=1}^{m}(y_j^i-1)/4m$。这样，从平均策略中获得的最终分数可以有效地消除由主观性造成的偏差。为了验证这一点，本文把 17 个标注分成几组，然后计算每对分组之间的 PCC。如表 2 所示，随着每组的标注数量的增加，PCC 也得到了改善。而当把所有标注分成两组时，它们的 PCC 达到了 0.940。这一现象说明，通过对更多标注者的标注进行平均，可以有效减少主观性。除了这个计算出来的复杂度分数，本文还将提供每个样本的标签分布，供研究者探索 IC 的更多特性。

3.3 数据集属性

• **标注一致性** 由于人类感知的主观性，不同人对视觉属性的感知可能是不同的，但在多次试验的分布下会保持稳定。在这里，本文试图证明多个标注者的主观标注的可靠性。按照常用的标准 [55], [56]，本文计算每对标注之间的 PCC 相关系数、Spearman 相关系数和 Kendall's tau 相关系数，并评估它

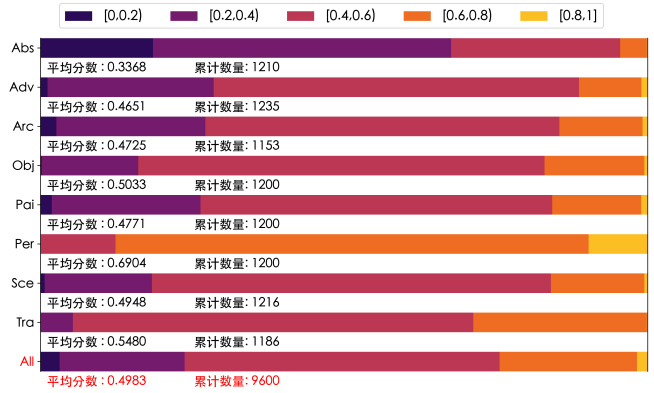

图 3. IC9600 数据集的标注分布。本文将 0 到 1 的分数平均划分为五个区间，并计算每个类别（用三个首字母缩写表示）的分数在这些区间内的比率。分布情况显示在堆积图中。每一行的下面是对应的平均分数和样本数量。整个数据集的信息显示在最后一行。

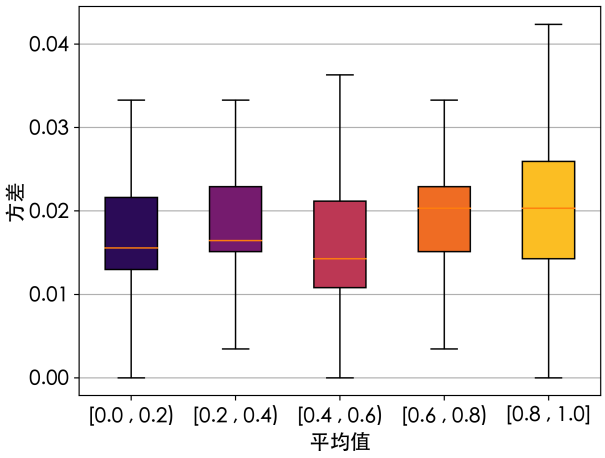

图 4. 方差的分布。所有样本根据其平均复杂度值分成五个区间。为了使箱线图表示更加清晰，这里省略了少数离群值。

们的统计属性。结果显示，平均 Pearson 相关系数、Spearman 相关系数和 Kendall's tau 相关系数分别为 0.54、0.53 和 0.48，在显著性为 0.01 时，所有配对的 P 值都小于 0.01，这表明了标注者之间的一致性。此外，与图像质量评估和图像美学评估的众包评估研究类似 [54], [57]，本文计算了标注的类内相关系数（intra-class correlation coefficient, ICC）。ICC 是一个最广泛使用的指标，用于衡量标注者之间的可靠性。ICC 数值越大，表明大部分的差异可以由图像的差异来解释，而不是源于标注者的个体差异，因此表明标注者之间有高度的一致性。本文使用了与 [54], [57] 相同的 ICC 模型。实验显示 ICC 为 0.518，优于 [54]（0.46）和 [57]（0.403）的结果。这也证明了本文标注的可靠性和一致性。

• **分布和标注**　所有图像按 7:3 的比例随机分为训练集和测试集。对于八个语义类别中的每一个，标注分数的分布显示在图 3 中。本文提出的 IC9600 数据集的类别分布是相对平衡的，每个类别包含约 1,200 幅图像。然而，不同语义类别的复杂度分布彼此不同。对于抽象类别，分数低于 0.2 的图像数量比其他类别都要多，这就导致了该类别拥有最低的平均分数。这也是合理的，因为抽象图像的内容大多是简单的几何图形。也可以注意到，人群类别的图像来自 WiderPerson [51]，它是一个密集的行人数据集，图片中的人群以一种拥挤和无序的方式呈现，使得图像非常复杂，因此人群类别的平均得分比其他类别高得多，而且没有一个得分低于 0.4。这个解释同样也适用于高平均分的交通类别。如图 3 最后一行所示，近一半的图像被分配到中等分数（即 $[0.4, 0.6)$），整个数据集呈现出以 0.5 左右为中心的对称高斯分布，这反映了真实世界的 IC 分布。此外，本文计算每个样本标注的平均值和方差，并根据其平均值将其分成五个区间。每个区间的方差分布都显示在图 4 的箱线图中。可以观察到，复杂度得分在 0.8 和 1.0 之间的样本方差相对较高，这表明这一区域的 IC 对人类来说可能略难明确区分。尽管如此，大多数的方差都低于 0.04，这证明了本文的数据集标注一致性很高。本文希望这样一个多样化的数据集可以帮助研究人员探索图像的更通用的 IC 属性，以及 IC 在各种计算机视觉任务中更广泛的应用。

4　预测图像复杂度

为了为今后的研究设定一个基线并验证所提数据集的合理性，本文设计了一个基础模型，称之为 ICNet。如图 5 所示，。细节分支捕捉空间和细节的表征，全局分支则编码背景和高级信息。然后本文将这两种特征连接起来，分别送到后面两个预测头中，用于 IC 热度图和 IC 分数的预测。此外，每个中间特征都由一个空间分布注意力模块来细化，该模块是专门为 IC 特征缩放而设计的，可以产生更有效的表征。

4.1　双分支提取器

IC 是一种基本的图像属性，它取决于整个图像中的低级表征和高级语义信息。这表明，所设计的模型需要同时具备从图像中挖掘出这两种特征的能力。给定一个深度卷积神经网络，Zeiler 等人 [59] 将特征激活投射回输入图像的像素中。可视化结果直观地证明来自 CNN 浅层的特征通常被简单的模式激活，如边缘、角落和角度等等。相反，深层的激活基本上是由更抽象的语义信息，如狗的脸决定的。基于上述关于 CNN 的经验现象，本文提出了一个用于特征提取的双分支网络。

本文模型的两个分支都由 ResNet18 [58] 修改而来。更具体地说，本文将 ResNet18 分为四个阶段。第一阶段将图像下采样为 1/4 分辨率的特征图，后面三个阶段的下采样系数分别为 1/8、1/16 和 1/32。ResNet18 末端的自适应平均池化层和全连接层都被丢弃。ResNet18 的所有阶段都是在 ImageNet [60] 数据集上预训练的，所以初始模型有很强的能力来捕捉通用的图像特征。对于全局分支，它包括所有四个阶段。并且输入到它的图像是低分辨率的（256×256）。通过如此操作，与小的输入尺寸相比，这种更深的网络产生了相

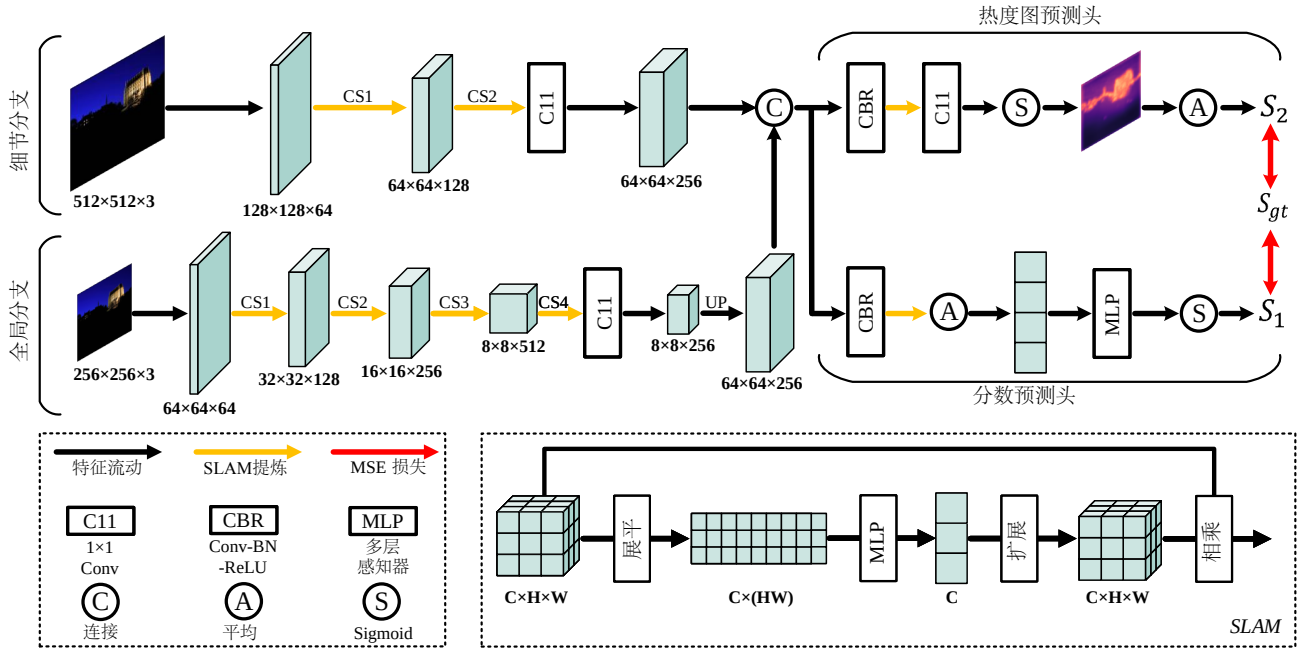

图 5. **本文提出的模型的结构图。该模型由从 ResNet18 [58] 修改得到的浅层提取器和深层提取器组成。箭头上的"CS(N)" 代表 ResNet18 的第 N 个卷积阶段。细节分支从高分辨率的图像中捕捉空间和低层次的特征，而全局分支则从较小的图像中提取全局和高层次的表征。之后这两种特征被连接起来，送到后面两个预测头中，用于热度图预测和分数预测。此外，后面有黄色箭头的特征图是由本文提出的空间分布注意力模块（SLAM）提炼之后得到的，它可以帮助特征根据空间分布进行缩放，为 IC 评估产生更有效的特征表示。**

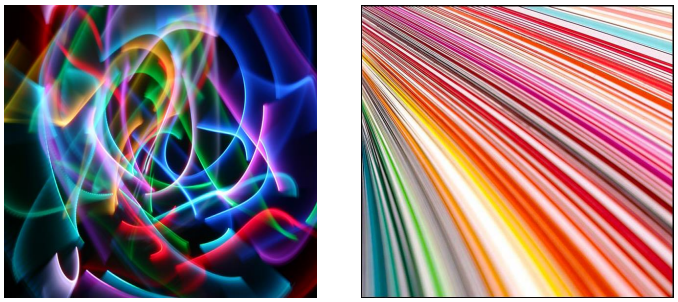

图 6. **从本文的数据集的抽象类别中挑选的两张图片。每张图片下的数字是真实的复杂度得分。尽管右边的图像包含了更多的颜色和线条，但左边的图像由于其中的线条布局不规则且无序，所以被标注更高的分数。**

对较大的感知视野，因此它可以对背景和高层次的抽象特征表示进行编码。相比之下，细节分支捕捉详细的空间信息，因此它只使用了 ResNet18 的前两个阶段，并且输入一个大尺寸的图像（即 512×512）。该分支的下采样系数仅为 1/8，可以产生大小为 64×64 的高分辨率特征图。

为了合并两个分支各自产生的特征图，本文在通道维度上采取了连接操作。注意到两个串联的特征图的空间维度需要相同，这意味着应该对全局分支的特征图进行上采样，或者对细节分支的特征图进行下采样。在这种情况下，细节特征所编码的空间信息在下面的章节中被证明是 IC 的一个重要因素。因此，本文对全局特征图进行了上采样，并将两种特征图连接起来，在保留空间细节的同时，充分地利用了低级和高级特征。

4.2 空间分布注意力模块

一般来说，具有更多纹理、边缘、物体等的图像被认为更复杂。一个常见的做法是将提取的特征平均池化为一个向量，并将其输入到多层感知器（multi-layer perceptron, MLP）中，最后再用一个 sigmoid 激活函数来预测复杂度得分。在这种方式下，该向量只编码图像中每个特征的平均强度。它缺少了一个重要的因素，即特征的空间布局，而后者在很大程度上决定了 IC 的强度。如图 6 所示，两张包含真实分数的图片是从本文数据集的抽象类别中挑选出来的。很明显，右边的图像有更多的颜色和线条，但左边的图像标注的 IC 分数更高。两幅图像的主要区别是图像元素的空间布局不同，这使得 IC 的感知具有差异。更具体地说，左边的图像是由弯曲和无序的线条形成的，而右边的图像比较均匀和一致。因此，标注者倾向于给左边的图像打高分。

受上述观察的启发，本文提出了一个注意力模块，它可以通过挖掘空间布局信息来自适应地调整激活强度，所以被称为空间分布注意力模块（spatial layout attention module, SLAM）。如图 5 所示，给定一个特征图 $\boldsymbol{F} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$，本文首先将它根据空间维度展平为二维特征图 $\boldsymbol{F_{fl}} \in \mathbb{R}^{\mathrm{C}\times(\mathrm{HW})}$。对于任何索引为 i 的单一通道，向量 $\boldsymbol{F_{fl}^{i}} \in \mathbb{R}^{\mathrm{HW}}$ 编码了 i_{th} 通道的特征在每个空间位置的激活。为了学习特征的布局如何影响 IC，本文将 $\boldsymbol{F_{fl}}$ 输入到 MLP 中，得到一个向量 $\boldsymbol{s} \in \mathbb{R}^{\mathrm{C}}$。这个操作与通道维度无关，这意味着对于通道维度 i 的索引，

输入是 $\boldsymbol{F}_{\boldsymbol{fl}}^{\boldsymbol{i}}$，而输出是标量 s^i。MLP 层由两个线性层组成，每个层后面都有一个激活函数，可以表示为：

$$s^i = \sigma_1(\boldsymbol{W_1} \times \sigma_0(\boldsymbol{W_0} \times \boldsymbol{F}_{\boldsymbol{fl}}^{\boldsymbol{i}} + \boldsymbol{b_0}) + \boldsymbol{b_1}), \tag{1}$$

这里 σ_0 和 σ_1 分别是 ReLU 和 sigmoid 激活函数，$\boldsymbol{W_0} \in \mathbb{R}^{(\mathrm{HW})\times 512}$ 和 $\boldsymbol{W_1} \in \mathbb{R}^{512\times 1}$ 是两个线性映射的权重，其中 $\boldsymbol{b_0} \in \mathbb{R}^{512}$ 和 $\boldsymbol{b_1} \in \mathbb{R}^1$ 是偏置参数。由于空间布局与不同的特征种类无关，因此 MLP 的权重和偏差通过通道维度共享。

通过(1)，MLP 可以灵活地评估 $\boldsymbol{F}$ 的第 i 层特征的空间布局对 IC 的影响程度（即 s^i）。如果这个特征的布局是有规律的、统一的，那么应该抑制这个特征。相比之下，当布局无序和混乱时，这一特征就应该被放大。为了满足这一要求，本文在空间维度上将 $\boldsymbol{s}$ 扩展到与 $\boldsymbol{F}$ 相同的大小，得到 $\boldsymbol{S}$。然后在这两个三维的张量上进行简单的元素相乘，得到最终的输出 $\boldsymbol{O}$。本文提出的 SLAM 的整体操作可以表示为：

$$\boldsymbol{O} = \boldsymbol{F} \cdot o_3(o_2(o_1(\boldsymbol{F}))), \tag{2}$$

其中 o_1、o_2、o_3 分别表示展平、MLP 和扩展算子。

如图 5 所示，本文在部分特征图后面添加 SLAM，这些特征图后面用黄色的箭头标识。另外，本文将大小在 32×32 以上的特征图降采样为 32×32，以降低计算成本。

4.3 预测复杂度分数和热度图

本文探讨了两种类型的 IC 模态。一个是描述图像整体复杂度的 IC 得分，另一个是描述局部区域复杂度的 IC 热度图。

为了产生这两种 IC 模态，本文将连接的特征图输入到后面两个预测头，即热度图预测头和分数预测头中。为了融合和平衡不同层次的特征，本文在两个预测头的入口处都设置了一个 Conv-BN-ReLU 块。之后，本文使用 SLAM 根据每个通道的空间布局，对其特征进行缩放。为了预测全局 IC 得分 S_1，本文将 SLAM 提炼的特征图送到全局平均池化层，得到一个特征向量作为之后的 MLP 层和 sigmoid 函数的输入。

对于热度图预测头，来自 SLAM 的特征图先被输入到一个 1×1 的卷积层，得到结果后再输入到 sigmoid 函数中。该操作将该特征图映射为一个单通道特征图，也就是复杂度热度图。然而这里出现了一个问题，即与分割任务类似，本文需要一个真实的 IC 热度图来评估用于反向传播的像素级回归损失。但实际上很难对这样的 IC 热度图进行标注。为了克服这一困难，本文提出了一种简单的弱监督方法，即只从真实 IC 得分 S_{gt} 中学习复杂度热度图。本文将生成的复杂度图进行平均，得到一个标量 S_2，然后计算它与 S_{gt} 之间的距离。通过该方式，预测头可以隐含地学习如何去预测局部 IC 强度，从而使平均值和真实值之间的距离最小。

在训练过程中，本文优化预测分数和真实标签之间的距离，可由均方误差（MSE）计算损失：

$$\mathcal{L}_1 = \frac{1}{N}\sum_{j=1}^{N}(S_1 - S_{gt})^2, \tag{3}$$

表 3
与 10 个传统方法和 4 个基于深度的方法的比较（在本文提出的数据集上训练和测试）。↑ (↓) 分别代表更高（更低）的更好。带有上标 U 的方法表示它们是无监督的方法。表中的'N/A' 表示该指标不适用于 UAE 方法。

	方法	指标			
		PCC ↑	SRCC ↑	RMSE ↓	RMAE ↓
传统方法	IC^{U} [61]	-0.006	0.053	0.343	0.552
	CR^{U} [13]	0.228	0.314	0.196	0.405
	FC^{U} [20]	0.459	0.439	0.342	0.558
	EN^{U} [20]	0.479	0.458	0.385	0.600
	SE^{U} [20]	0.534	0.498	0.136	0.327
	NR^{U} [20]	0.556	0.541	0.188	0.394
	ED^{U} [15]	0.569	0.491	0.226	0.427
	AR^{U} [62]	0.571	0.481	0.234	0.445
	HOG+SVR [63]	0.689	0.689	0.118	0.299
	SIFT+SVR [64]	0.885	0.861	0.069	0.242
深度方法	UAE^{U} [22]	0.651	0.635	N/A	N/A
	SAE [22]	0.865	0.860	0.074	0.240
	AlexNet [60]	0.924	0.920	0.064	0.222
	ResNet18 [58]	0.935	0.928	0.061	0.222
	HyperIQA [24]	0.935	0.935	0.067	0.229
	P2P-FM [56]	0.940	0.936	0.056	0.208
	ICNet (本文)	**0.949**	**0.945**	**0.053**	**0.205**

$$\mathcal{L}_2 = \frac{1}{N}\sum_{j=1}^{N}(S_2 - S_{gt})^2, \tag{4}$$

其中 N 是一个批次的样本总数，总体损失由 $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 计算得到：

$$\mathcal{L} = \lambda \times \mathcal{L}_1 + (1-\lambda) \times \mathcal{L}_2, \tag{5}$$

其中 λ 控制两部分损失的比重。

5 实验和结果

5.1 实验设置

• **实现细节** 本文基于 PyTorch [65] 框架，使用两个 NVIDIA GTX 1080TI GPU 来实现所提出的模型。本文使用 ImageNet [60] 数据集上的预训练模型来初始化双分支提取器的每个阶段的参数。批量（批量大小为 64）随机梯度下降法（SGD）被用来优化模型。其中动量被设定为 0.9，权重衰减设置为 0.001。本文将初始学习率设定为 0.05，每 10 个 epoch 之后除以 5。总体训练 30 个 epoch 的时间约为 1 小时。该模型的训练使用的是上面提出的默认训练-测试划分。所有的训练图像都采用随机水平翻转的方式进行增强。此外，本文将 λ 设置为 0.9，以获得两种 IC 模态之间的性能平衡。对于评价指标的计算，本文使用 S_1 作为最终的预测分数。

• **评价指标** 本文使用 Pearson 相关系数（PCC）[20]、Spearman 相关系数（SRCC）[23]、根均方误差（root mean square error, RMSE）和根均绝对误差（root mean absolute error, RMAE）来评价这些方法。

PCC 的定义为:

$$\rho(X,Y)=\frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sqrt{\sum_{i=1}^{N}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{N}(Y_i-\mu_Y)^2}}, \quad (6)$$

其中 X 和 Y 代表预测分数和相应的真实主观分数。μ_X 和 μ_Y 是 X 和 Y 的平均值。N 是总的图像数量。SRCC 的计算为:

$$\rho'=1-6\frac{\sum_{i=1}^{N}(r_i-r_i')^2}{N^3-N}, \quad (7)$$

其中，r_i 和 r_i' 代表预测分数和真实分数按降序排列时，第 i 个样本的排名。此外，RMSE 的计算方法是:

$$r(X,Y)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-Y_i)^2}, \quad (8)$$

RMAE 表示为:

$$m(X,Y)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}|X_i-Y_i|}. \quad (9)$$

• **对比方法**　本文将所提出的方法与几种传统方法和基于深度的方法进行比较。传统方法包括 IC (image colorfulness) [61], CR (compression ratio) (JPEG format) [13], FC (feature congestion) [20], EN (entropy) [20], SE (subband entropy) [20], NR (number of regions) [20], ED (edge density) [15], AR (auto-regressive) model [62], HOG+SVR (support vector regression with histogram of orientated gradients) [63], [66], SIFT+SVR (support vector regression with scale-invariant feature transform) [64]。深度方法包含经典模型（即 AlexNet [60] 和 ResNet18 [58]）。此外，本文与其他两种用于图像质量评估的方法（HyperIQA [24] 和 P2P-FM [56]）进行比较。在 [22] 中提出的基于 CNN 的 IC 评估方法也被加入到本文的比较中，包括无监督的（UAE）方法和有监督的（SAE）方法。UAE 方法将来自预训练模型的中间激活层（如 ReLU）的特征图平均到一个单一激活分数来代表图像复杂度。由于平均激活只能代表同一类别图像内的相对分数，所以本文只采用 PCC 和 SRCC 来评估这种方法。类似地，SAE 方法也是首先从预先训练好的 CNN 的中间层提取特征，其中的区别在于，这些特征随后被送到岭回归模型，并在真实 IC 分数的监督下进行训练。此外，根据 [22]，与其他架构如 ResNet、DenseNet、EfficientNet 等相比，VggNet 可以得到最佳性能。因此本文的复现实验是在 VggNet 上进行的。由于 [22] 没有提供关于他们提取的是哪一层特征的细节，本文选择产生最佳性能的特征来报告结果。本文使用默认的训练-测试划分，在 IC9600 上训练和测试这些方法。对于没有发布代码的算法，本文根据他们的论文来复现代码。

5.2　实验结果

本文在表 3 中展示了不同方法的评估结果，并得到以下结论。首先，基于手工特征的方法很难胜过基于深度学习的方法。它们中的大多数都有低于 0.6 的 PCC，而深度方法均超过 0.9。可以观察到表 3 第 1 行的 IC 的 PCC 非常接近 0，揭示了图像中颜色的变化可能不是图像复杂度的关键因素。其他的方法如 CR、FC、NR 等等，都显示出与复杂度有更好的相关性。注意到它们每一个都只是 IC 的单一组成部分，它们可以从特定角度来评估 IC，但并不是完整衡量这一抽象概念的综合指标。在这些传统方法中，手工的 SIFT 特征与 SVR 的组合产生了最好的性能。这表明具有提取不同尺度的空间特征能力的 SIFT 与 IC 有很好的相关性，但由于只是一个局部特征描述，所以它的表现受到了限制。

第二，在本文的大规模和高质量数据集的支持下，CNN 可以提取高级的表征，这些表征比低级的特征更能模拟人类对 IC 的感知。这个结论可以由表 3 中深度方法的高性能证明。即使是简单地从分类网络（AlexNet 和 ResNet）修改而来的普通 CNN，其表现也远远超过了传统方法。

第三，本文的模型在四个指标上超过了所有的比较方法，证明了本文模型的优越性。HyperIQA 和 P2P-FM 的表现比本文提出的方法略差，因为这些专门为图像质量评估设计的方法忽略了一些影响人类对 IC 感知的因素，比如说细节和全局信息。本文提出的模型可以避免这个问题，并同时从细节和全局的角度对 IC 进行建模，从而取得良好效果。注意到虽然 UAE 和 SAE 的方法都显示出与本文方法的巨大性能差距，但结果是合理的。首先，UAE 方法没有采用任何人类评价信息，而是纯粹依赖于复杂区域的激活通常较高的假设。尽管这一假设可能部分正确，但 UAE 产生的 IC 分数是整个激活图的平均值，它没有考虑到图像的元素布局，而这一点正是

表 4
与 10 个传统方法和 4 个基于深度的方法的比较（在本文提出的数据集上训练它们，并在小规模的 SAVOIAS 数据集 [22] 上测试它们）。

	方法	指标			
		PCC ↑	SRCC ↑	RMSE ↓	RMAE ↓
传统方法	IC^U [20]	0.230	0.243	0.290	0.485
	CR^U [13]	0.271	0.305	0.257	0.452
	FC^U [20]	0.430	0.456	0.259	0.454
	EN^U [20]	0.448	0.466	0.375	0.567
	SE^U [20]	0.352	0.345	0.261	0.454
	NR^U [20]	0.580	0.595	0.244	0.438
	ED^U [15]	0.467	0.449	0.273	0.460
	AR^U [62]	0.497	0.485	0.261	0.454
	HOG+SVR [63]	0.380	0.350	0.253	0.447
	SIFT+SVR [64]	0.704	0.695	0.185	0.382
深度方法	UAE^U [22]	0.763	0.763	N/A	N/A
	SAE [22]	0.750	0.750	0.189	0.394
	AlexNet [60]	0.819	0.818	0.183	0.387
	ResNet18 [58]	0.843	0.845	0.177	0.380
	HyperIQA [24]	0.826	0.831	0.184	0.390
	P2P-FM [56]	0.836	0.842	0.179	0.383
	ICNet (本文)	**0.866**	**0.868**	**0.176**	**0.379**

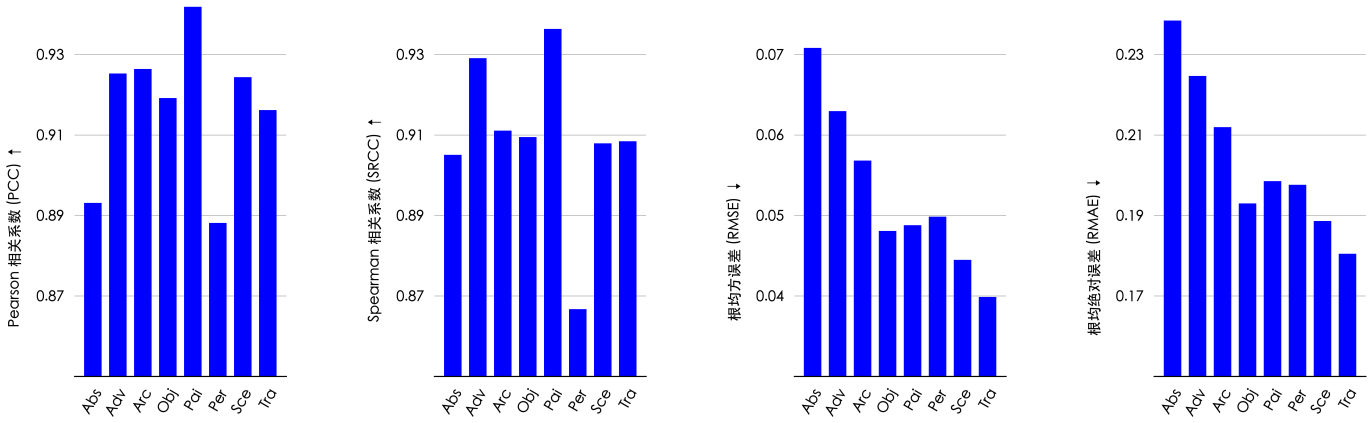

图 7. 本文所提数据集 IC9600 中的每个语义类别在 PCC、SRCC、RMSE 和 RMAE 等指标上的性能。

被本文的实验所证明的 IC 评估的关键，因此该方法很难挖掘出高层次的 IC 因素。对于 SAE 方法，由于 SAVOIAS 数据集 [22] 中缺乏训练数据，作者只采用预训练的 CNN 进行特征提取，并采用岭回归进行分数预测，其中 CNN 的参数在整个过程中是固定的，这限制了其潜在的学习能力。

注意对于无监督方法内部的比较或它们与有监督方法之间的比较，评估相对 IC 的 PCC 和 SRCC 比 RMAE 和 RMSE 更有代表性，因为不同方法的 IC 强度校准是不同的。然而，对于 PCC 和 SRCC，可以从表 3 中观察到，无监督的方法通常不如有监督的方法，这也是合理的，因为本文的大规模和良好标注的数据集可以为有监督的方法带来巨大优势，使得它们可以逐步捕捉更广泛的 IC 因素，这比那些从有限角度预先定义的特征更加有效和完整，从而产生更好的性能。

此外，为了验证本文模型的泛化能力，还在最近的 SAVOIAS [22] 数据集上对比了不同方法的实验结果，可见表 4。注意到这个数据集是小规模的，而且真实的分数是为每个类别单独标注的（总共有七个类别）。因此，本文在 IC9600 所有样本上训练每个方法，然后在 SAVOIAS 上测试这些方法，并报告七个类别的平均结果。结果显示，本文提出的 ICNet 也能超过所比较的方法，这证明了本文的模型具有显著的泛化能力。

本文还提供了 ICNet 在本文所提数据集的每个语义类别上的评价结果，如图 7 所示。可以看到，抽象类别的性能在每个指标上都相对较低。我们猜测这可能来自于抽象类别图像中内容的多样性，使得模型很难预测出一致的结果。此外，在八个类别中，人群类别的 SRCC 明显最低，而在 RMSE 和 RMAE 方面，它的表现相对较好。这是合理的，因为图 3 中显示的人群的复杂度大多分布在一个高复杂度区间，即 $[0.6, 0.8)$。因此，预测一个具体的分数会比较容易，但明确预测每张图像的复杂度排序比其他类别更难，导致 SRCC 最低。相反，图 3 中绘画图像的分布在五个复杂度区间内比较均匀，因此它获得了最佳的 SRCC 性能。

为了更好地理解 IC，本文在图 8 中展示了一些由本文的模型预测的可视化结果。可以发现，预测的分数与真实的分数非常接近。其中每一对图的右边是由细节分支预测的复杂度热度图。对于该可视化，本文通过双线性插值对热度图进行上采样，使其达到图像的大小，然后将其与输入图像进行 α 混合。从这些热度图中观察到，本文的模型可以精确地找到图像中的视觉复杂区域。还可以发现，复杂区域大多集中在有大量物体、纹理、边缘、变化等的位置，这些区域使人很难明确描述。这些热度图在给模型（机器）提供理解图像中复杂度分布的指导方面有很大的潜力，并可能在未来被应用于各种任务，如图像裁剪、自动驾驶、图像生成、广告设计等。图中最后一行显示了两种失败案例。第一个和第二个样本有很多局部纹理，所以模型倾向于预测一个更高的分数。第三个例子的大部分区域是空白的，因此模型预测的复杂度得分较低。收集更多此类样本的图像可能有助于在未来解决这些失败案例。此外，本文在图 9 中绘制了热度图的训练动态。可以观察到，在早期，由于初始模型对 IC 的直接感知有限，所以复杂度热度图在整个图像上的分布不管其内容如何都是均匀的。随着训练的进行，该模型在真实复杂度的监督下，需要从平均化的局部区域预测全局的复杂度，这间接地需要对每个像素进行细粒度的预测。因此热度图中显现了围绕前景物体的高复杂度区域，而低复杂度和高复杂度像素之间的边界也从粗糙到平滑逐渐细化。最后，从整个有数千张图像的数据集中进行学习，模型逐渐学会合理的复杂度热度图的构建模式，从而使平均激活和真实得分之间的损失最小，因此预测的复

表 5
两个分支和 SLAM 的消融分析。

设置	指标			
	PCC ↑	SRCC ↑	RMSE ↓	RMAE ↓
只有细节分支	0.929	0.927	0.063	0.222
只有全局分支	0.939	0.933	0.061	0.219
双分支	0.944	0.939	0.058	0.215
双分支 + SE [67]	0.944	0.939	0.059	0.218
双分支 + CBAM [68]	0.943	0.942	0.059	0.217
双分支 + BAM [69]	0.944	0.938	0.057	0.214
双分支 + SLAM	**0.949**	**0.945**	**0.053**	**0.205**

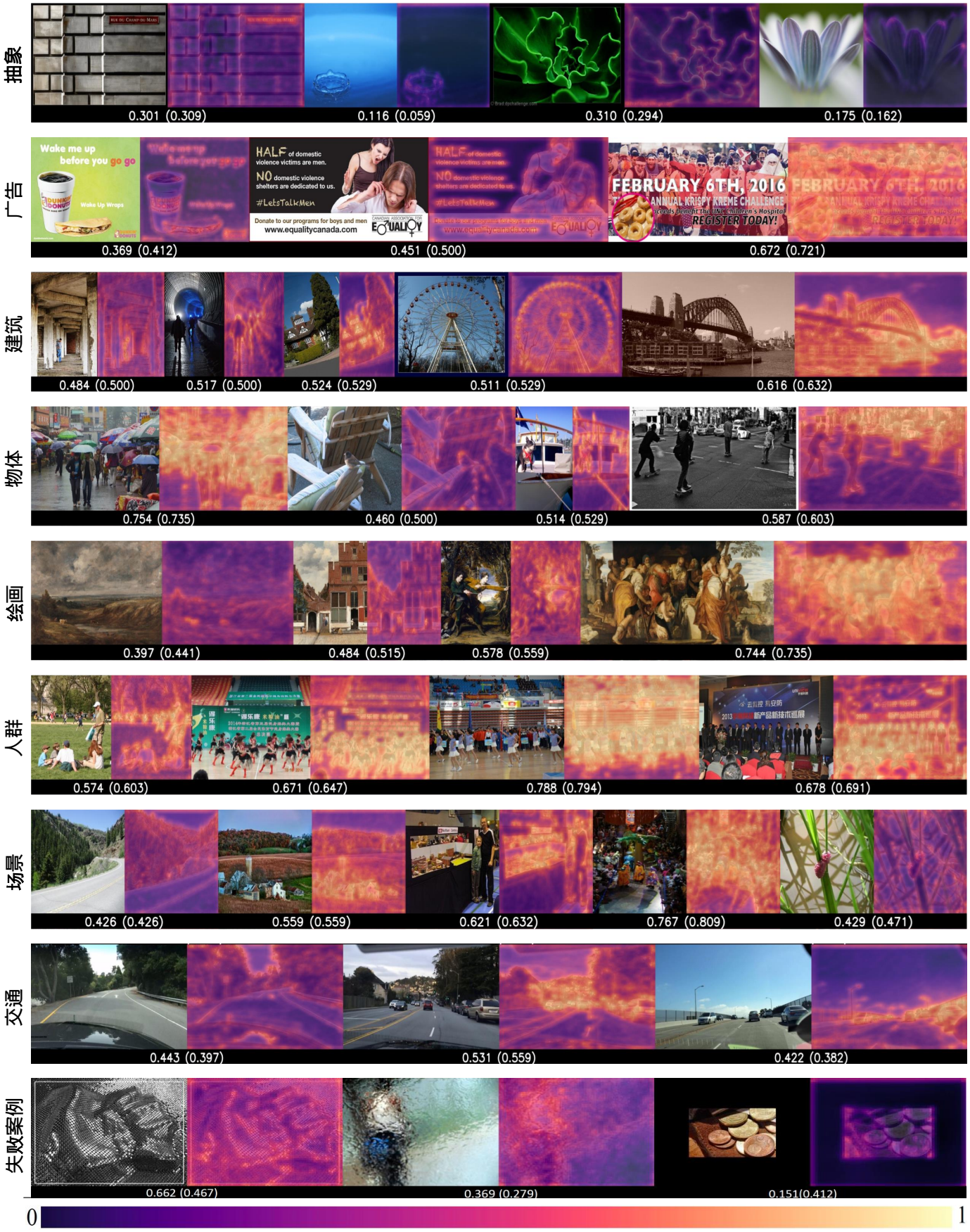

图 8. 可视化结果。在每一对图像中，左边（右边）的图像分别是输入图像（预测的复杂度热度图）。括号内的数字代表本文模型预测的复杂度分数，而括号外的数字是由标注者标注的真实分数（归一化为 $0-1$）。最后一行显示了几个失败的案例，它们预测的分数与真实的分数相差较大。

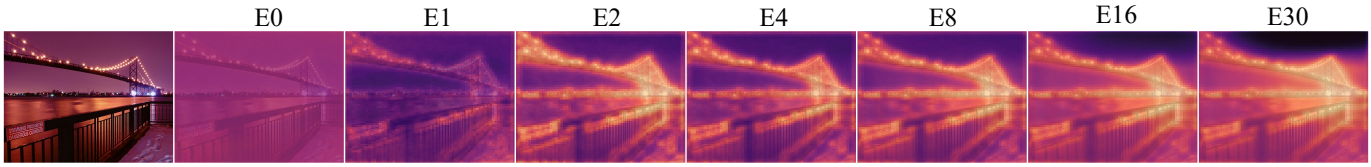

图 9. 训练过程中复杂度热度图的动态变化，‘E’ 表示 ‘Epoch’。初始情况下，模型在第一个 epoch (E0) 很难为每个像素预测合理的 IC。经过几个 epoch 的训练，具有复杂纹理和不规则布局的局部区域显示出较高的激活值，而干净平坦的背景（如天空）则被赋予较低的 IC 分数。在接下来的训练中，整个 IC 热度图被逐渐细化，变得更加精细，并在简单和复杂区域之间呈现出明显区别。

杂度热度图可以精确反映整个图像的像素级复杂度。

5.3 消融分析

本文研究了 ICNet 每个组成部分的有效性，结果见于表 5。当只使用细节分支时，得到的 PCC 为 0.929，比只使用全局分支 (PCC 为 0.939) 低。尽管细节分支输入的图像分辨率更高，但它主要是捕捉低层次和空间信息，而全局分支可以从小尺寸的图像中提取抽象和高层次的表征。结果表明，全局信息在 IC 评估中更为关键。尽管如此，它们都是人们感知 IC 的重要因素。因此，当把两个分支结合起来时，性能可以提高到 0.944 (PCC)。此外，当在中间层插入所提出的 SLAM 时，由于考虑到了特征的空间分布信息来细化特征，因此 PCC 进一步提高到 0.949。本文还通过使用 SE [67]、CBAM [68] 和 BAM [69] 分别取代 SLAM，研究了其他注意机制的效果，其性能表现与只使用全局和细节分支接近。我们推测，这些注意机制是为通用的特征提取而设计的，但没有考虑到 IC 的内在特性，因此在 IC 评估任务中的泛化性不理想。本文提出的 SLAM 专门为 IC 评估设计，它利用了空间布局信息，所以能进一步提高性能。

6 应用和讨论

本节探索 IC 在多种计算机视觉任务中的应用方法。我们展示了使用 IC 的三种方式及在六个下游任务中的应用。还讨论了图像复杂度的更广泛的潜在应用，表明 IC 也适用于计算机视觉或图像处理之外的许多其他领域。

6.1 作为一个辅助任务

多任务学习是一种普遍而直观的想法，它对深度学习模型的可靠改进已经在许多有影响力的工作中得到了证明 [70], [71]。其主要思想是增加更多与主要任务相关的监督信息，并同时对其进行优化，以帮助模型学习更鲁棒的特征，提高泛化能力。在这里，本文将复杂度评估作为一项辅助任务，并使用本文训练好的 ICNet 自动生成监督信号。本文在四个视觉任务上验证了增加辅助复杂度评估子分支的有效性。

• **图像美学评估**　以前的工作已经证明 IC 是美学评估中的一个重要依据 [72]–[74]。他们发现，复杂度对美学的评价有负面影响。受此启发，本文试图通过与 IC 协同优化来提高图像

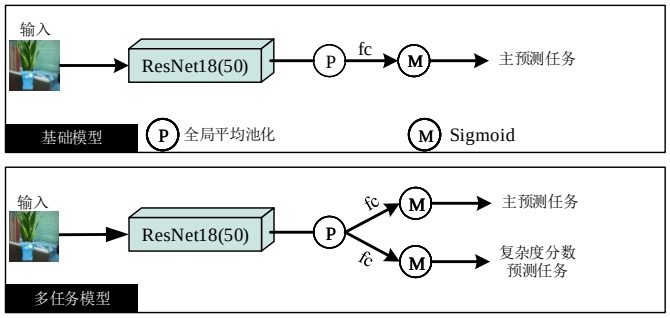

图 10. 图像质量评估、图像美学评估和图像分类等任务的原始模型和多任务模型的对比。

美学评估的性能。具体来说，在本文的实验中，将基础模型设置为一个普通的 ResNet18 网络，通过改变最后一个全连接层来输出美学分数或美学等级，而多任务模型是在基础模型上进行修改，即在最后一个平均池化层后面增加了另一个分支来预测 IC 得分。为了训练多任务模型，本文使用在所提数据集上训练完成的 ICNet，生成每张图像的真实复杂度分数。本文在 AADB [23] 和 CUHKPQ [75] 数据集上使用相同的实验设置训练和测试基础模型和多任务模型。基础模型和多任务模型的流程图可见图 10。实验结果显示在表 6 的第 1 行。可以观察到，AADB 和 CUHKPQ 数据集的 PCC 和 SRCC 都提高了 1−2 个点，这证明了 IC 可以藉由辅助分支的方式提高图像美学评估的性能。

• **图像质量评估**　图像质量评估也是一项主观任务，与人类的感知密切相关。IC 的强度与图像质量感知也有很大的联系。例如，当注入高频噪声时，图像的复杂度会增加，而当引入低频模糊时，复杂度则会下降 [62]。因此，本文也将 IC 应用于 LIVEC [35] 和 KADID [38] 数据集，以多任务训练方式进行图像质量评估。实验设置与上述图像美学评估的实验类似，同样使用 ResNet18 骨干网络。基础模型和多任务模型的流程图显示在图 10 中。实验结果显示在表 6 的第 2 行。与图像美学评估类似，当协同优化这两个任务时，骨干网络被强迫提取更多的通用和鲁棒的特征，以便能够同时满足这两个任务的需求。因此，当引入另一个 IC 预测分支时，IQA 的性能也可以得到改善。

• **图像分类**　常见的图像分类任务也可能从 IC 的监督中受

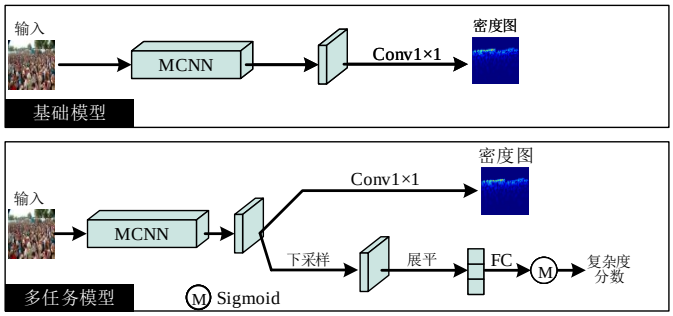

图 11. 基础模型（MCNN）和多任务框架在人群计数任务中的对比。

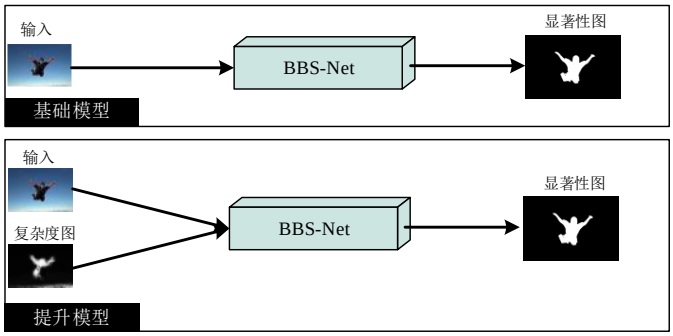

图 12. 基础模型（BBS-Net）和改进后的模型在显著物体检测任务中的对比。

益。由于 IC 有时是不同类别的图像的判别特征，学习挖掘有效的 IC 评估特征有助于分类任务。本文使用 ResNet50 在 Tsinghua Dogs [76] 数据集上进行了实验。如图 10 所示，本文在 ResNet50 的末尾添加了一个子分支，以回归 IC 得分，并反向传播 MSE 误差，该误差衡量预测得分和真实得分之间的距离。采用额外 IC 信息的改进结果显示在表 6 的第 5 行中。

• **人群计数**　人群计数旨在从单个图像中估计人的数量。一般来说，图像中人越多，图片就越复杂。因此，决定图像复杂度的特征也可以被用于人群计数。对于这项任务，本文使用 MCNN [25] 进行实验。MCNN 由三个平行的 CNN 组成，使用这些 CNN 的合并特征图来预测人群密度图（任务 1）。本文通过在合并的特征图后面插入另一个预测头（任务 2）来改进 MCNN。这个头部包含一个下采样层，先将特征图缩小，随后将特征图展平，最后使用一个全连接层预测复杂度得分。通过这种方式，该模型可以由两个任务协同优化。在这里，真实的复杂度得分是由 ICNet 产生的。多任务模型使用与基础模型相同的训练设置。训练原始 MCNN 和多任务 MCNN 的流程图显示在图 11。训练框架和进行评估的 WorldExpo10 [77] 和 UCF QNRF [78] 数据集是从 [79] 修改而来的。表 6 第 3 行的结果显示，并行的复杂度预测任务可以减少人群计数的误差（MAE 和 RMSE 较低）。

6.2　作为一个新模态

本文提出的模型 ICNet 可以从热度图预测头生成像素级的复杂度图。这个精细的复杂度图可以被视为一种新模态，为模型理解图像的局部复杂程度提供指导。

• **显著性检测**　IC 对显著物体检测有着至关重要的影响 [26]。可以从图 8 中观察到，具有高复杂度的区域通常在很大程度上与显著的物体重叠。另外，对于视觉上简单的图像，通常很容易将前景物体从背景区域中分离出来，而对于复杂的图像，似乎很难分割出显著的物体。为了验证 IC 信息是否能够帮助找到显著的物体，本文在 DUTS [80] 和 PASCAL-S [81] 数据集上做了实验。本文选择 BBS-Net [85] 作为基础模型。它使用两个分支从 RGB 模态和深度模态中寻找显著物体。由于 BBS-Net 提供了一种高效的、通用的多模态提取和融合策略。原有的深度通道可以简单地被其他模态所取代。对于基础模型，本文使深度分支的输入为零。为了利用复杂度模态，本文让深度分支的输入为本文的 ICNet 生成的复杂度热度图。然后，使用相同的设置来训练和测试这两个模型。这个任务的流程图显示在图 12 中。表 6 显示，与基础模型相比，具有复杂度模态输入的模型具有较高的 max F-measure 和较低的 MAE。它证明了 IC 模态可以通过为模型提供图像中简单和复杂区域的先验指导来帮助分割显著的物体。

6.3　作为一个先验权重

直观地说，视觉上复杂的图像可能难以识别或分割。为了解决这个问题，本文为复杂图像或复杂局部区域设置高权重，使模型更加关注它们。

• **图像分类**　在这里，本文试图将一幅图像的 IC 分数作为先验知识，以代表模型正确分类的困难程度。由于图像分类是一项图像级别的任务，本文利用图像级别的复杂度得分对每个样本进行加权。本文使用 ResNet50 [58] 对 Food-101 数据集 [82] 进行了实验。具体来说，对于基础模型，本文使用交叉熵损失来优化模型，损失定义为 $\ell_{ce} = -(\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j \ln p_i^j)/N$，其中 N 是总图像数，C 是总类别。如果 j 是真实标签，那么 $y_i^j = 1$,，否则 $y_i^j = 0$。p_i^j 是最后的 softmax 层的输出。而对于改进后的模型，本文将损失修改为 $\ell'_{ce} = -(\sum_{i=1}^{N} w_i \sum_{j=1}^{C} y_i^j \ln p_i^j)/N$，其中 w_i 代表第 i 个样本的权重，即提出的 ICNet 预测的复杂度得分。通过这样，复杂的图像将被赋予更高的损失权重来进行优化。表 6 第 5 行的结果表明，可以通过利用先验的 IC 权重有效地提高分类性能。

• **图像分割**　对于图像分割，本文在计算损失时，根据 ICNet 生成的 IC 热度图，给每张图像的像素加权（即高复杂度的区域被赋予更高的权重）。本文在 PASCAL VOC 2012 [83] 和 CityScapes [84] 数据集上进行了实验。采用具有代表性的 Deeplabv3 分割模型 [86] 进行比较。对于基础模型，图像的像素级交叉熵损失掩码 $\boldsymbol{L}$ 被定义为 $\boldsymbol{L} = -\sum_{j=1}^{C} \boldsymbol{G_j} \odot \ln \boldsymbol{Y_j}$，其中 C 是总的类别，$\boldsymbol{G_j}$ 和 $\boldsymbol{Y_j}$ 是第 j 个类别的真实分割图和预测图，$\odot$ 表示元素级别的乘积。本文将掩码的平均损失 $\boldsymbol{L}$ 反向传播，表示为 $\ell = \frac{1}{H}\frac{1}{W}\sum_{i=1}^{H}\sum_{j=1}^{W} \boldsymbol{L}_{ij}$，其中 H 和 W

表 6
IC 提高了各种视觉任务的性能。'maxF' 和 'Acc' 代表 max F-measure 和准确度。'本文'= 基线方法 + IC 信息。

任务	数据集	指标	基线	本文	数据集	指标	基线	本文
图像美学评估	AADB [23]	PCC ↑ SRCC ↑	0.702 0.693	**0.713** **0.705**	CUHKPQ [75]	ACC↑	0.856	**0.878**
图像质量评估	LIVEC [35]	PCC ↑ SRCC ↑	0.842 0.806	**0.851** **0.818**	KADID [38]	PCC ↑ SRCC ↑	0.706 0.724	**0.730** **0.745**
人群计数	WorldExpo10 [77]	MAE ↓ RMSE ↓	19.33 28.64	**16.83** **25.04**	UCF-QNRF [78]	MAE↓ RMSE↓	276 425	**250** **386**
显著物体检测	DUTS [80]	maxF ↑ MAE ↓	0.855 0.043	**0.865** **0.039**	PASCAL-S [81]	maxF ↑ MAE ↓	0.893 0.071	**0.899** **0.068**
图像分类	Food-101 [82]	ACC↑	0.814	**0.834**	Tsinghua Dogs [76]	ACC↑	0.804	**0.818**
图像分割	VOC2012 [83]	mIoU ↑ pixACC ↑	0.594 0.858	**0.610** **0.872**	CityScapes [84]	mIoU ↑ pixACC ↑	0.612 0.923	**0.623** **0.925**

是掩码的高度和宽度。对于改进后的模型，损失掩码 $\boldsymbol{L}'$ 的计算方法是：$\boldsymbol{L}' = -\boldsymbol{W} \odot \sum_{j=1}^{C} \boldsymbol{G_j} \odot \ln \boldsymbol{Y_j}$，其中 $\boldsymbol{W}$ 代表由 ICNet 产生的复杂度图。表 6 的最后一行显示，使用 IC 热度图可以帮助提高图像分割的性能。

6.4 计算机视觉领域外的应用

除了本文上面提到的方式，图像复杂度还可以应用于更广泛的领域。本文从大量的相关工作中调研和总结了潜在的应用。

• **心理学** 理解视觉复杂度是研究人类感知的一个重要媒介。格式塔心理学 [87] 起源于发现感官输入和感知复杂度之间的联系，并将视觉复杂度作为揭示人脑如何感知的基础。该领域的研究从单一的视觉形式、视觉阵列到视觉显示 [88]，大多都可以归结为对视觉复杂度内部机制的挖掘和理解。此外，视觉复杂度已被证明会影响多种心理学领域，包括注意力和情绪系统 [89]、视觉模式编码机制 [90] 和视觉记忆性 [22] 等等。

• **艺术** 视觉复杂度在很大程度上决定了对艺术作品的评价。例如，在 [91] 中发现，复杂度和建筑外观之间存在强烈的正向线性关系。Gartus 等人 [92] 提出，复杂度主要在数量和结构的维度影响抽象模式。此外，素描、绘画、摄影等美学与视觉复杂度之间的关系仍在广泛的研究之中 [5], [6], [93], [94]。因此，一个准确的 IC 评估工具在帮助评估艺术作品的内在价值方面将使广泛的艺术领域从业者受益。

• **网页和广告设计** 大量的研究证明，复杂度在网页和广告设计中起着关键作用。Pieters 等人 [95] 发现，广告中密集的复杂视觉特性会伤害顾客的注意力和对品牌的态度。类似地，直播中过度复杂的背景也被证明对个人的购买意向有负面影响 [96]。在用移动设备购物时，复杂度在用户的满意度方面显示出更大的影响 [97]。此外，简单和清晰已经成为现代网页设计的趋势，因为较低的复杂度被证明有更强的吸引力 [7], [98]。上述研究意味着在很多商业活动中应该谨慎控制视觉复杂度。

• **讨论** 上面列出的应用大多是在计算机视觉社区之外的，而且跨越了广泛的领域。其中，对自动和可靠的 IC 评估方法存在迫切需求，因为：首先，大多数研究人员通过采用边缘检测或图像压缩等传统方法来评估 IC，这些方法不能准确反映图像的综合复杂度，因此可能导致结论有偏差。其次，由于没有可靠的 IC 评估工具，在一些研究中，图像的复杂度大多是通过多个人的评分来收集，这种方法不灵活，成本高，耗时长，因此阻碍了大规模的 IC 应用。为了解决上述问题，我们相信本文高质量的数据集和可靠的 IC 评估模型将为推动进一步的 IC 研究和潜在应用提供支持。

7 总结

本文深入探索了图像复杂度评估这一具有挑战性且长期被忽视的问题。首先通过建立一个大规模的基线数据集来解决最关键的数据缺失问题，该数据集由来自不同类别的 9600 张精心标注的图像组成。基于这个数据集，本文随后提供了一个基线模型，称为 ICNet，用来评估图像的复杂度得分，它可以与人类的感知达到很高的匹配度。此外，本文还尝试将复杂度评价模型应用于六个任务，实验结果表明，IC 可以帮助提高它们的性能。我们希望所提出的数据集、模型和应用探索能够鼓励和促进 IC 的进一步研究。

参考文献

[1] T. Feng, Y. Zhai, J. Yang, J. Liang, D. Fan, J. Zhang, L. Shao, and D. Tao, "IC9600: A benchmark dataset for automatic image complexity assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[2] A. Forsythe, "Visual complexity: is that all there is?" in *International Conference on Engineering Psychology and Cognitive Ergonomics*, 2009.

[3] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity." *Journal of Experimental Psychology: Human Learning and Memory*, vol. 6, no. 2, pp. 174–215, 1980.

[4] C. Heaps and S. Handel, "Similarity and features of natural textures." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, no. 2, pp. 299–320, 1999.

[5] L. Sun, T. Yamasaki, and K. Aizawa, "Relationship between visual complexity and aesthetics: application to beauty prediction of photos," in *European Conference on Computer Vision*, 2014.

[6] J. McCormack and A. Lomas, "Deep learning of individual aesthetics," *Neural Computing and Applications*, vol. 33, no. 1, pp. 3–17, 2021.

[7] A. N. Tuch, J. A. Bargas-Avila, K. Opwis, and F. H. Wilhelm, "Visual complexity of websites: Effects on users' experience, physiology, performance, and memory," *International Journal of Human-Computer Studies*, vol. 67, no. 9, pp. 703–715, 2009.

[8] F. Meng, H. Li, K. N. Ngan, L. Zeng, and Q. Wu, "Feature adaptive co-segmentation by complexity awareness," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4809–4824, 2013.

[9] R. Grover, D. K. Yadav, D. Chauhan, and S. Kamya, "Adaptive steganography via image complexity analysis using 3d color texture feature," in *International Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity*, 2018.

[10] M. Li and M. Bai, "A mixed edge based text detection method by applying image complexity analysis," in *World Congress on Intelligent Control and Automation*, 2012.

[11] C. Yanyan, W. Huijuan, and M. Xinjiang, "Digital image enhancement method based on image complexity," *International Journal of Hybrid Information Technology*, vol. 9, no. 6, pp. 395–402, 2016.

[12] P. Li, Y. Yang, W. Zhao, and M. Zhang, "Evaluation of image fire detection algorithms based on image complexity," *Fire Safety Journal*, vol. 121, pp. 103 306–103 317, 2021.

[13] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal, "Computerized measures of visual complexity," *Acta Psychologica*, vol. 160, pp. 43–57, 2015.

[14] L. Dai, K. Zhang, X. S. Zheng, R. R. Martin, Y. Li, and J. Yu, "Visual complexity of shapes: a hierarchical perceptual learning model," *The Visual Computer*, 2021.

[15] X. Guo, Y. Qian, L. Li, and A. Asano, "Assessment model for perceived visual complexity of painting images," *Knowledge-Based Systems*, vol. 159, pp. 110–119, 2018.

[16] Y.-Q. Chen, J. Duan, Y. Zhu, X.-F. Qian, and B. Xiao, "Research on the image complexity based on neural network," in *International Conference on Machine Learning and Cybernetics*, 2015.

[17] A. Olivia, M. L. Mack, M. Shrestha, and A. Peeper, "Identifying the perceptual dimensions of visual complexity of scenes," in *The Annual Meeting of the Cognitive Science Society*, vol. 26, no. 26, 2004, pp. 1041–1044.

[18] A. M. Iliyasu, A. K. Al-Asmari, M. A. AbdelWahab, A. S. Salama, M. A. Al-Qodah, A. R. Khan, P. Q. Le, and F. Yan, "Mining visual complexity of images based on an enhanced feature space representation," in *IEEE International Symposium on Intelligent Signal Processing*, 2013.

[19] A. Miniukovich and A. De Angeli, "Quantification of interface visual complexity," in *International Working Conference on Advanced Visual Interfaces*, 2014.

[20] S. E. Corchs, G. Ciocca, E. Bricolo, and F. Gasparini, "Predicting complexity perception of real world images," *PloS one*, vol. 11, no. 6, pp. 1–22, 2016.

[21] Z. B. Fan, Y.-N. Li, J. Yu, and K. Zhang, "Visual complexity of chinese ink paintings," in *ACM Symposium on Applied Perception*, 2017.

[22] E. Saraee, M. Jalal, and M. Betke, "Visual complexity analysis using deep intermediate-layer features," *Computer Vision and Image Understanding*, vol. 195, pp. 102 949–102 968, 2020.

[23] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *IEEE International Conference on Computer Vision*, 2017.

[24] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[25] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[26] M. Liu, K. Gu, G. Zhai, and P. Le Callet, "Visual saliency detection via image complexity feature," in *IEEE International Conference on Image Processing*, 2016.

[27] S. Arthur, "Entropy, visual diversity, and preference," *The Journal of General Psychology*, vol. 129, no. 3, pp. 300–320, 2002.

[28] N. Gauvrit, F. Soler-Toscano, and H. Zenil, "Natural scene statistics mediate the perception of image complexity," *Visual Cognition*, vol. 22, no. 8, pp. 1084–1091, 2014.

[29] H. C. Purchase, E. Freeman, and J. Hamer, "Predicting visual complexity," in *International Conference on Appearance*, 2012.

[30] M. P. Da Silva, V. Courboulay, and P. Estraillier, "Image complexity measure based on visual attention," in *IEEE International Conference on Image Processing*, 2011.

[31] R. Rosenholtz, Y. Li, and L. Nakano, "Measuring visual clutter," *Journal of Vision*, vol. 7, no. 2, pp. 17–17, 2007.

[32] X. Guo, T. Kurita, C. M. Asano, and A. Asano, "Visual complexity assessment of painting images," in *IEEE International Conference on Image Processing*, 2013.

[33] H. Yu and S. Winkler, "Image complexity and spatial information," in *International Workshop on Quality of Multimedia Experience*, 2013.

[34] M. A. Abdelwahab, A. M. Iliyasu, and A. S. Salama, "Leveraging the potency of cnn for efficient assessment of visual complexity of images," in *International Conference on Image Processing Theory, Tools and Applications*, 2019.

[35] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[36] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler, and C. Redies, "Jenaesthetics subjective dataset: analyzing paintings by subjective scores," in *European Conference on Computer Vision*, 2014.

[37] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[38] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *International Conference on Quality of Multimedia Experience*, 2019.

[39] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *IEEE International Conference on Computer Vision*, 2011.

[40] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale

database for aesthetic visual analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[41] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[42] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, 2018.

[43] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[44] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2815–2826, 2019.

[45] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2018.

[46] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang, "Group maximum differentiation competition: Model comparison with few samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 851–864, 2018.

[47] Z. Wang and K. Ma, "Active fine-tuning from gmad examples improves blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[48] K. Sheng, W. Dong, M. Chai, G. Wang, P. Zhou, F. Huang, B.-G. Hu, R. Ji, and C. Ma, "Revisiting image aesthetic assessment via self-supervised feature learning," in *AAAI Conference on Artificial Intelligence*, 2020.

[49] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka, "Automatic understanding of image and video advertisements," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014.

[51] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "Widerperson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 380–393, 2019.

[52] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[53] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, 2018.

[54] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[55] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision*, 2016.

[56] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[57] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1338–1350, 2016.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[59] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 2012.

[61] E. Alghamdi, E. Velloso, and P. Gruba, "Auvana: An automated video analysis tool for visual complexity," *OSF Preprints*, 2021.

[62] K. Gu, J. Zhou, J.-F. Qiao, G. Zhai, W. Lin, and A. C. Bovik, "No-reference quality assessment of screen content pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4005–4018, 2017.

[63] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[64] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999.

[65] B. Steiner, Z. DeVito, S. Chintala, S. Gross, A. Paszke, F. Massa, A. Lerer, G. Chanan, Z. Lin, E. Yang *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019.

[66] Z. Xie, J. Liu, C. Liu, Y. Zuo, and X. Chen, "Optical and sar image registration using complexity analysis and binary descriptor in suburban areas," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[67] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[68] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision*, 2018.

[69] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[70] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *International Joint Conference on Artificial Intelligence*, 2016.

[71] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[72] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *British Journal of Psychology*, vol. 95, no. 4, pp. 489–508, 2004.

[73] R. Reber, N. Schwarz, and P. Winkielman, "Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?" *Personality and Social Psychology Review*, vol. 8, no. 4, pp. 364–382, 2004.

[74] K. N. Ochsner, "Are affective events richly recollected or simply familiar? the experience and process of recognizing feelings past." *Journal of Experimental Psychology: General*, vol. 129, no. 2, p. 242, 2000.

[75] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[76] D.-N. Zou, S.-H. Zhang, T.-J. Mu, and M. Zhang, "A new dataset of dog breed images and a benchmark for finegrained classification," *Computational Visual Media*, vol. 6, no. 4, pp. 477–487, 2020.

[77] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[78] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *European Conference on Computer Vision*, 2018.

[79] J. Gao, W. Lin, B. Zhao, D. Wang, C. Gao, and J. Wen, "C^3 framework: An open-source pytorch code for crowd counting," *arXiv preprint arXiv:1907.02724*, 2019.

[80] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[81] Y. Xu, D. Xu, X. Hong, W. Ouyang, R. Ji, M. Xu, and G. Zhao, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *IEEE International Conference on Computer Vision*, 2019.

[82] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.

[83] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[84] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[85] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *European Conference on Computer Vision*, 2020.

[86] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[87] W. Köhler, "Gestalt psychology," *Psychologische Forschung*, vol. 31, no. 1, pp. XVIII–XXX, 1967.

[88] D. C. Donderi, "Visual complexity: a review." *Psychological Bulletin*, vol. 132, no. 1, p. 73, 2006.

[89] N. Sadeh and E. Verona, "Visual complexity attenuates emotional processing in psychopathy: Implications for fear-potentiated startle deficits," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 12, no. 2, pp. 346–360, 2012.

[90] S. F. Chipman and M. J. Mendelson, "Influence of six types of visual structure on complexity judgments in children and adults." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 5, no. 2, p. 365, 1979.

[91] D. Hussein, "A user preference modelling method for the assessment of visual complexity in building façade," *Smart and Sustainable Built Environment*, vol. 9, no. 4, pp. 483–501, 2020.

[92] A. Gartus and H. Leder, "Predicting perceived visual complexity of abstract patterns using computational measures: The influence of mirror symmetry on complexity perception," *PloS one*, vol. 12, no. 11, pp. 1–22, 2017.

[93] A. Forsythe, M. Nadal, N. Sheehy, C. J. Cela-Conde, and M. Sawey, "Predicting beauty: Fractal dimension and visual complexity in art," *British Journal of Psychology*, vol. 102, no. 1, pp. 49–70, 2011.

[94] A. Tuch, S. Kreibig, S. Roth, J. Bargas-Avila, K. Opwis, and F. Wilhelm, "The role of visual complexity in affective reactions to webpages: Subjective, eye movement, and cardiovascular responses," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 230–236, 2011.

[95] R. Pieters, M. Wedel, and R. Batra, "The stopping power of advertising: Measures and effects of visual complexity," *Journal of Marketing*, vol. 74, no. 5, pp. 48–60, 2010.

[96] X. Tong, Y. Chen, S. Zhou, and S. Yang, "How background visual complexity influences purchase intention in live streaming: The mediating role of emotion and the moderating role of gender," *Journal of Retailing and Consumer Services*, vol. 67, p. 103031, 2022.

[97] S. Sohn, B. Seegebarth, and M. Moritz, "The impact of perceived visual complexity of mobile online shops on user's satisfaction," *Psychology & Marketing*, vol. 34, no. 2, pp. 195–214, 2017.

[98] A. Krishen, "Perceived versus actual complexity for websites: their relationship to consumer satisfaction," *The Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, vol. 21, 2008.