

从单幅图像估计景深的模型到底学到了什么

胡立华^{1,2)}, 朵安鹏^{1,2)}, 杨海峰¹⁾, 张继福¹⁾, 胡占义^{2)*}

¹⁾ (太原科技大学计算机科学与技术学院 太原 030024)

²⁾ (中国科学院自动化研究所模式识别国家重点实验室 北京 100190)
(huzy@nlpr.ia.ac.cn)

摘要: 目前, 基于深度学习的单幅图像景深估计取得了显著的进展, 在一些公开的室内外数据集上均取得了非常高的估计精度. 然而, 不论是基于监督学习还是基于自监督学习的景深估计模型, 这些模型到底学习到了图像的什么性质, 使得其能够对景深进行很好的估计呢? 为此, 从 2 个侧面对这个问题进行了量化的测试和分析. 对于“无纹理”区域, 通过模型对这些区域与该区域邻域估计的景深之间的关系, 探究了这些区域的估计景深是不是其邻域估计景深的某种“填充”效应; 其次分析了模型对高视觉显著性区域估计的景深是否具有更高的估计精度. 测试结果显示, 无纹理区域与其邻域的估计景深分布确实存在比较高的相似性, 但当前景深估计模型的估计精度和图像视觉显著性的关联性不是很强. 所得结果对“景深估计模型解析”“景深估计模型改进”等相关工作均具有一定的参考价值, 例如, 今后在设计和训练景深估计模型的工作中, 有必要充分考虑输入图像的视觉显著性效应, 从而提高模型对高视觉显著性区域的景深估计精度, 以便更好地服务下游任务.

关键词: 景深估计; 模型解析; 无纹理区域; 视觉显著性

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20110

What Have the Single Image Based Depth Prediction Models Learnt

Hu Lihua^{1,2)}, Duo Anpeng^{1,2)}, Yang Haifeng¹⁾, Zhang Jifu¹⁾, and Hu Zhanyi^{2)*}

¹⁾ (College of Computer Sciences and Technology, Taiyuan University of Science and Technology, Taiyuan 030024)

²⁾ (National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

Abstract: Recently, single-image based depth learning via deep learning has achieved tremendous progress, and impressive prediction accuracy has been reported on both indoor and outdoor benchmark datasets. However, what have been learnt by such models from single images under either supervised or self-supervised learning framework? It seems this fundamental problem is rarely discussed in the literature up to now. In this work, this problem is investigated from the following two aspects: at first, for those texture-poor or no texture regions, it is tested whether the corresponding predicted depths are somewhat filling-in effect of the depths in their close neighborhood region. Secondly, it is assessed whether the regions with high visual saliency usually have better depth prediction performance. Our test results show that the predicted depths in texture-poor regions indeed have high correlation with the depths in their close neighborhood region. However, the accuracy of the estimated depth is not particularly related to the visual saliency of the input image. The above results could be of reference value for both model analysis and model design, for example, visual saliency of input image could be taken into account in the model design and

收稿日期: 2022-12-29; 修回日期: 2024-03-14. 基金项目: 国家自然科学基金(61991423, 62273248). 胡立华(1982—), 女, 博士, 教授, 硕士生导师, CCF 会员, 主要研究方向为计算机视觉、人工智能与模式识别; 朵安鹏(1999—), 男, 硕士研究生, 主要研究方向为计算机视觉; 杨海峰(1980—), 男, 博士, 教授, 博士生导师, 主要研究方向为计算机视觉、机器学习; 张继福(1963—), 男, 博士, 教授, 博士生导师, 主要研究方向为人工智能与模式识别; 胡占义(1961—), 男, 博士, 研究员, 博士生导师, 论文通信作者, 主要研究方向为计算机视觉.

training to enhance the prediction accuracy of high saliency regions, so as to better serve the down-stream vision tasks.

Key words: depth estimation; model analysis; texture-poor region; visual saliency

从几何的观点看,从单幅图像无法恢复某个像素对应的空间点的景深信息,因为该像素对应的投影射线上的所有点都投影为该图像点,是一个典型的“一对多”映射问题.但近年来,鉴于深度网络强大的图像表达能力和自然图像结构分布的非随机性特性,很多理论上的“非适定性”问题在数据驱动下具有了“可行性”,包括从单幅图像估计景深的问题.经过近几年的深入研究,目前从单幅图像估计景深已是一项相对比较成熟的技术^[1-6].基于单幅图像的景深估计模型,不论是在有监督还是自监督框架下训练的模型,它们在公开的室内外数据集上均取得了非常高的景深估计精度.那么,模型到底学习到了什么,使其从图像能很好地估计每个像素对应的景深?包括无纹理或弱纹理图像区域对应的景深?这个问题现有文献中却鲜有报道.

关于基于单幅图像的景深估计模型到底学习到了什么这一问题,2019年,van Dijk等^[7]发现,在以驾驶场景为主的KITTI^[8]数据集上,目前的景深估计模型中“障碍物与地面的接触区域”是一个关键特征,该特征的有/无和区域的大小直接决定对应障碍物的景深能否被可靠估计;Hu等^[9]发现,对于KITTI数据集,目前的模型更关注图像中道路远处的区域.这一结论与van Dijk等^[7]的结论有些相悖.Hu等^[9]发现,对于以室内场景为主的NYU Depth V2^[10]数据集,物体封闭区域和边缘处的像素对景深估计具有更大的贡献.因此,当前“景深估计模型”在很大程度上仍是一个“黑盒子”,模型的可解释性有待进一步分析和揭示.

基于单幅图像的景深估计模型可以估计像素级的景深,这意味着景深估计模型对弱纹理甚至无纹理的图像区域也能估计对应的景深.弱纹理和无纹理区域的景深估计问题,是诸如立体视觉等基于多幅图像之间特征匹配的三维重建方法无法克服的困难.因为对无纹理或弱纹理区域,无法可靠地提取图像特征,所以无法可靠地进行特征匹配和三维重建.那么,基于单幅图像的景深估计模型对无纹理区域估计的景深与对该区域邻域内估计的景深是否存在联系,特别是能给出一些定

量刻画,目前尚未见到相关报道.

图像是“场景语义”的表现形式.那么,基于单幅图像的景深估计模型所估计的景深图,其精度与场景语义有什么关系?场景语义是一个比较泛的概念,很难给出确切的定义和度量.鉴于目前文献中已经报道了大量关于估计“视觉显著性图”的模型^[11-13],本文也将探索图像的视觉显著性与景深估计精度的关系这个问题.

1 问题描述和方法途径

本文将从2个方面对基于单幅图像的景深估计模型的性质进行分析和测试:

(1) 模型对图像无纹理区域估计的景深是否为其邻域区域估计景深的某种“填充”效应?

(2) 模型对高视觉显著性区域估计的景深是否比对低视觉显著性区域估计的景深有更高的精度?

1.1 图像无纹理区域与其邻域区域估计景深之间的关系

需要考虑的问题是:如何描述无纹理区域.本文采用等量随机黑白像素充填不同尺寸方框的形式(下文称之为“测试框”)表示无纹理区域,使用随机黑白像素填充是为了避免“黑色像素填充”或“白色像素填充”可能会产生的内容歧义.如黑色填充区域会与室内场景中的“电视机”“显示屏”等物体或者室外场景中“车辆”等物体混淆,白色填充区域会与室内场景中的“白色毛巾”等物体或室外场景中“天空”等区域混淆.

根据测试图像的归一化显著性图区分图像高(低)显著性区域.通常,图像角点的密集度反映图像内容的复杂程度,因此把图像中角点密集的区域称为高复杂度区域,角点稀疏的区域称为低复杂度区域.

影响测试结果的因素有:测试框的大小、测试框邻域的大小、测试框处图像的复杂度和测试框处图像的视觉显著度等.为此,采用如下的途径进行测试:

(1) 改变测试框的大小,保持其邻域像素的个数尽量与测试框的像素个数相近,分析测试框内

与其邻域内像素的估计景深分布的相似性;

(2) 将测试框置于图像高复杂度和低复杂度区域时, 分析测试框与对应邻域内的估计景深分布的相似性;

(3) 将测试框置于图像高视觉显著性和低视觉显著性区域时, 分析测试框与对应邻域内的估计景深分布的相似性.

考虑测试框内有多个像素, 且测试框和邻域内的像素数量不等, 本文采用重叠指数^[14-15]度量 2 个分布之间的相似性, 因为重叠指数与 DL-散度不同, 不存在非对称性问题; 同时, 还统计分析了测试框与对应邻域内像素的景深估计误差的相似性.

1.2 图像不同视觉显著性区域的景深估计误差之间的关系

鉴于视觉显著性可以表现人眼在观察某个场景的过程中对于不同区域的关注程度, 因此本文测试了图像的高视觉显著性区域的景深估计误差是否低于低视觉显著性区域的景深估计误差.

对于一幅图像, 首先计算其视觉显著性图, 然后根据显著性阈值将图像分为高显著性区域和低显著性区域, 再计算不同区域中像素的景深估计误差, 最后比较高显著性区域与低显著性区域内像素的景深估计误差之间的差异.

2 测试模型、评估方法和测试数据

2.1 景深估计模型

为了使得到的结论尽量与具体使用的测试模型无关, 本文选取近年来文献中 3 种具有代表性的“监督模型”和 3 种具有代表性的“自监督模型”进行测试. 其中, 基于有监督学习的景深估计模型包括 Lee 等^[16]提出的 BTS(from big to small), Yuan 等^[17]提出的 NeW CRFs 和 Bhat 等^[18]提出的 AdaBins; 基于自监督学习的景深估计模型包括 Li 等^[19]提出的 StructDepth, Peng 等^[20]提出的 EPCDepth(excavating the potential capacity of self-supervised monocular depth estimation)和 Sun 等^[21]提出的 SC-DepthV3.

下面对这些景深估计模型进行简要介绍:

(1) 基于有监督学习的景深估计模型

BTS 模型由用于密集特征提取的编码器和用于景深估计的解码器 2 部分组成. 在编码器中, BTS 引入膨胀空间金字塔池化(atrous spatial pyramid pooling, ASPP)模块, 它通过在不同采样率下对输入的特征图进行多尺度池化, 并在不同尺度上对特征图和上下文信息进行编码, 提高网

络的感受野和语义表达能力. 在解码器中, BTS 引入多尺度局部平面引导(local planar guidance, LPG)模块引导特征学习, 以更好地捕捉场景的几何信息, 最后将该模块在不同尺度下的估计结果进行融合得到精细的景深图.

NeW CRFs 是一种基于条件随机场(conditional random fields, CRFs)的景深估计模型, 在计算成本和性能方面取得了良好的平衡. 为了降低计算复杂度, NeW CRFs 采用基于窗口的方法, 在每个窗口内执行 FC-CRFs(fully-connected CRFs)优化; 此外, 该模型还利用多头注意力机制捕捉图像中节点之间的关系.

AdaBins 采用编码器-解码器卷积神经网络结构, 在此基础上引入基于 ViT(vision Transformer)的模块. AdaBins 将景深离散为多个区间, 并且根据输入的图像自适应地计算景深区间的中心值, 像素最终的估计景深是这些中心值的概率加权线性组合.

(2) 基于自监督学习的景深估计模型

StructDepth 模型由景深网络、曼哈顿法线检测模块和平面检测模块 3 部分构成. 其中, 景深网络从输入图像估计出景深图, 曼哈顿法线检测模块和平面检测模块用于生成监督信号以训练网络. 给定一幅输入图像, 模型可以输出像素级的景深图.

SC-DepthV3 模型引入外部训练的单目景深估计模型, 用于生成景深先验, 即伪景深, 并提出新的损失函数以提高自监督学习的性能; 此外, 该模型还采用残差连接、空洞卷积、分组卷积和通道注意力机制等技术, 提高网络的性能和稳定性.

EPCDepth 使用数据嫁接、自蒸馏损失和全尺度网络等技术, 在不引入额外约束条件的前提下充分挖掘自监督方法的潜力.

本文实验中使用的 BTS, NeW CRFs, AdaBins 和 SC-DepthV3 景深估计模型均在 NYU Depth V2 和 KITTI 数据集上训练; StructDepth 模型在 NYU Depth V2 数据集上训练; EPCDepth 模型在 KITTI 数据集上训练.

2.2 视觉显著性估计模型

眼动实验表明, 不同人在同一场景下有非常高的概率关注场景中的同一区域, 即高视觉显著性区域. 早期的视觉显著性方法主要基于手工特征和启发式途径, 目前这类方法已被更准确的基于深度学习的模型取代. 本文使用 SalGAN^[11], FastSal^[12]和上下文编码器-解码器网络(contextual encoder-decoder network, CEDN)^[13]这 3 种视觉显著性估计

模型均是基于深度学习的模型, 均在 SALICON 数据集^[22]上训练. SALICON 数据集是在视觉显著性估计任务中最常用的数据集之一, 其数据量比 MIT300 和 MIT1003 等大很多; 其中的图像是 MS-CoCo^[23]的子集. 研究人员根据人类眼动和鼠标轨迹之间的相关性^[22], 利用模型将鼠标轨迹转换为眼动数据, 得到数据集中的标注信息.

下面简要介绍这 3 种视觉显著性估计模型.

(1) SalGAN 将生成对抗网络(generative adversarial network, GAN)应用于视觉显著性估计任务. 该模型包含生成器和判别器 2 部分. 其中, 生成器用于处理输入图像, 并输出对应的显著性图; 判别器用于判断输入的显著性图是真实的还是由生成器生成的. 这 2 部分通过对抗训练的方式迭代, 最终实现对输入图像的显著性区域的估计.

(2) FastSal 基于编码器-解码器结构, 其中, 编码器用于提取图像的多尺度特征, 解码器用于将这些特征映射到显著性空间. 该模型在保证精度的同时还尽可能做到了轻量化, 模型参数量约为 2.5M.

(3) CEDN 基于编码器-解码器结构, 引入 ASPP 模块捕捉输入图像的多尺度信息, 并充分地考虑了上下文信息. 网络训练过程包含预训练和微调 2 个阶段: 在预训练阶段, 将解码器用于物体和场景分类任务进行训练, 学习图像的底层特征; 在微调阶段, 使用带标注的显著性数据进行有监督微调.

2.3 重叠指数

已有文献中有多种估计 2 个密度分布函数相似性程度的方法, 如 KL-散度, Kolmogorov-Smirnov 检验等. 鉴于重叠指数具有对称性并有开源的软件包 Overlapping^[14-15], 可以方便计算由 2 个离散样本集表示的分布之间的重叠度, 故本文采用重叠指数对无纹理区域与其邻域区域的估计景深分布的相似性进行度量, 以判定模型对无纹理区域估计的景深是否为其邻域区域估计景深的某种填充效应. 如果重叠指数大, 说明模型对无纹理区域估计的景深很可能是其有纹理的邻域区域中估计景深的某种填充效应; 否则, 说明在景深估计中, 邻域的估计景深对无纹理区域估计的景深没有明显的影响.

重叠指数的计算步骤如下. 对于 2 组离散数据 $x=(x_1, x_2, \dots, x_m)$ 和 $y=(y_1, y_2, \dots, y_n)$, 假定其对应的连续随机变量为 X 和 Y , 则它们对应的概率密度函数 f_X 和 f_Y 可以通过核密度估计(kernel density estimation, KDE)确定, 公式分别为

$$f_X(x) = m^{-1} \sum_{i=1}^m K\left(\frac{x-x_i}{h}\right),$$

$$f_Y(y) = n^{-1} \sum_{i=1}^n K\left(\frac{y-y_i}{h}\right).$$

其中, K 表示核函数, 软件包提供的可选核函数包括 Gaussian, rectangular 和 biweight 等; h 表示带宽, 其大小会影响核密度估计的准确性, 可以手动指定, 也可以选择某一策略自适应估计, 软件包提供的可选策略包括 bw.nrd0, bw.nrd, bw.ucv 和 bw.bcv 等, 本文测试中使用软件包提供的 bw.nrd0 进行带宽估计. 最后, 重叠指数可通过

$$\eta(X, Y) = \int_{\mathbb{R}^n} \min(f_X(z), f_Y(z)) dz.$$

计算. 重叠指数的值在 0~1, 当 x 和 y 表示的分布越相似, 重叠指数越接近于 1; 否则越接近于 0.

2.4 测试数据

为了避免测试结果过于依赖使用的测试数据, 本文在包含不同场景类型的数据上进行测试, 测试数据组成如表 1 所示. 室内场景测试数据由 NYU Depth V2 和 SALICON 数据集中的部分图像组成. SALICON 数据集主要用于视觉显著性估计, 图像的场景类型分布非常广泛, 由于目前的景深估计模型还很难泛化到如此广的场景, 因此不能直接将该数据集中的所有图像用于测试. 本文通过手工筛选, 从 SALICON 数据集中抽取了 250 幅室内场景的图像用于测试. 为了使 2 个数据集中用于测

表 1 测试数据的场景分布情况

数据集	场景类型	图像数量/幅
NYU Depth V2	office	14
	bedroom	62
	living room	44
	classroom	14
	kitchen	52
	dining room	19
	study	5
	bookstore	6
	bathroom	1
	home office	9
	playroom	9
	others	5
KITTI	city	75
	residential	80
	campus	25
	road	70
SALICON	indoors	250

试的图像数量尽量接近, 从 NYU Depth V2 数据集的测试集中, 本文根据不同场景类型的图像数量按比例随机抽取图像, 最终所得图像包含 16 类场景共计 250 幅. 由于其中一些场景的原始图像数量很少(仅有 1~5 幅), 因此在表 1 中把这些场景统一称为 others. NYU Depth V2 数据集中各类场景的景深范围为 0~10 m. 室外场景测试数据从 KITTI 数据集中随机抽取得到, 最终所得图像共 4 类场景共计 250 幅, 景深范围为 0~80 m.

3 测试框估计景深与其邻域估计景深的分布差异分析

当测试框在某一位置时的测试流程如图 1 所示, 具体测试步骤如下.

Step1. 在每幅测试图像 I_i 中确定不同类别的候选测试点, 候选测试点分别位于图像中的高视觉显著性区域、低视觉显著性区域、高复杂度区域和低复杂度区域.

Step2. 分别从每种类别的候选测试点中随机选取 8 个作为测试点, 因此每幅图像中共有 32 个测试点; 测试框和其邻域框以测试点为中心构建, 测试框边长为 11×11 , 15×15 和 21×21 , 对应的邻域框边长为 15×15 , 21×21 和 31×31 , 即每幅图像中共有 96 个测试框; 使用等量黑白像素在测试框内随机填充, 得到的填充图像记为 I'_i , 其中, $i=1, 2, \dots, 96$ 表示测试框的序号, 将 I'_i 送入某个单目景深估计模型, 得到景深图记为 D'_i .

Step3. 在景深图 D'_i 中, 对于 3 种不同尺寸的测试框, 分别不重复随机采样 80, 180 和 350 个像素; 测试框内和邻域的采样结果记为 x 和 y , 对应的连续随机变量为 X 和 Y , 因此可以得到它们之间的重叠指数 $\eta(X, Y)$; 考虑到测试框尺寸不同, 框内像素数量不同, 因此对边长为 11×11 , 15×15 和 21×21 的测试框, 分别随机采样 20, 50 和 100 次, 最后计算它们的均值和方差, 将均值视为测试框在此处的测试结果.

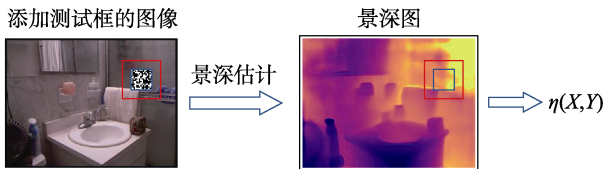


图 1 测试流程图

最后, 使用上述策略计算所有测试框的重叠指数, 并统计在不同视觉显著性估计模型、不同景

深估计模型和不同数据集下和不同类型的测试框重叠指数的均值和方差.

3.1 测试框位于图像高(低)复杂度区域时估计景深的分布差异分析

根据图像中 Harris 角点的密集度确定图像的不同复杂度区域. 将测试框置于某一位置, 若框内 Harris 角点像素的占比大于复杂度阈值 T_{texture} , 表示该测试框位于图像高复杂度区域; 否则, 表示测试框位于图像低复杂度区域.

对 6 种景深估计模型进行测试, 结果基本相同. 鉴于篇幅所限, 表 2 所示为 BTS 和 StructDepth 在 NYU Depth V2 和 KITTI 数据集的测试结果, 其余结果见补充资料^①. 可以看出, 测试框位于高复杂度区域时, 测试框与其邻域内像素的估计景深分布的重叠指数普遍大于测试框位于低复杂度区域时的结果, 并且随着测试框的增大, 重叠指数逐渐降低; 有监督模型 BTS 得到的重叠指数一般大于自监督模型 StructDepth 的结果; 在 NYU Depth V2 数据集上的重叠指数略低于在 KITTI 数据集上的结果. 测试结果表明, 不论在 NYU Depth V2 还是 KITTI 数据集上, 测试框位于图像中何种类型的区域以及使用有监督学习模型还是自监督学习模型, 测试框与其对应邻域的景深分布总具有较高的重叠度(重叠指数均大于 0.596 0).

3.2 测试框位于图像高(低)视觉显著性区域估计景深的分布差异分析

使用 SalGAN, FastSal 和 CEDN 这 3 种模型, 在 NYU Depth V2, SALICON 和 KITTI 数据集上估计对应的显著性图. 图像中像素对应的显著值若大于阈值 $\tilde{T}_{\text{high-saliency}}$, 则视为高显著性像素; 若低于阈值 $\tilde{T}_{\text{low-saliency}}$, 则视为低显著性像素. 将测试框置于某一位置, 若框内高显著性像素占比大于显著性阈值 T_{saliency} , 表示测试框位于高显著性区域; 否则, 表示测试框位于低显著性区域.

同样对 6 种景深估计模型进行测试, 其流程与第 3.1 节基本相同, 区别仅是将测试框置于不同显著性区域, 在不同模型下得到的结果基本相同. 鉴于篇幅有限, 表 3 所示为使用 SalGAN 模型估计显著性图, BTS 和 StructDepth 景深估计模型在 NYU Depth V2 和 KITTI 数据集下的结果, 其余测试结果见补充资料^①. 测试结果表明, 测试框位于

① <https://github.com/DK05310214/What-Have-the-Single-Image-Based-Depth-Prediction-Models-Learnt>

表 2 测试框位于图像高(低)复杂度区域时的重叠指数的均值和方差对比

景深估计模型	数据集	场景类型	测试框尺寸	均值	方差
BTS	NYU Depth V2	高复杂度	11	0.748 3	0.010 8
			15	0.708 9	0.014 1
			21	0.659 5	0.018 2
		低复杂度	11	0.752 0	0.012 3
			15	0.708 7	0.015 7
			21	0.650 4	0.020 8
	KITTI	高复杂度	11	0.729 9	0.009 5
			15	0.710 1	0.010 3
			21	0.693 8	0.011 0
		低复杂度	11	0.721 0	0.006 3
			15	0.689 6	0.007 2
			21	0.657 5	0.010 8
StructDepth	NYU Depth V2	高复杂度	11	0.706 6	0.006 5
			15	0.671 1	0.007 3
			21	0.611 3	0.008 7
		低复杂度	11	0.689 6	0.007 8
			15	0.650 5	0.008 2
			21	0.596 0	0.010 2
	KITTI	高复杂度	11	0.723 3	0.005 4
			15	0.696 7	0.005 7
			21	0.667 1	0.006 4
		低复杂度	11	0.706 1	0.006 8
			15	0.677 1	0.007 4
			21	0.645 2	0.008 1

表 3 测试框位于高(低)视觉显著性区域时的重叠指数的均值和方差

景深估计模型	数据集	场景类型	测试框尺寸	均值	方差
BTS	NYU Depth V2	高视觉显著性	11	0.761 9	0.011 5
			15	0.732 9	0.014 2
			21	0.685 9	0.018 5
		低视觉显著性	11	0.747 2	0.012 2
			15	0.705 7	0.015 8
			21	0.647 5	0.021 5
	KITTI	高视觉显著性	11	0.726 2	0.007 2
			15	0.696 3	0.008 8
			21	0.662 2	0.011 3
		低视觉显著性	11	0.722 8	0.005 9
			15	0.690 8	0.007 8
			21	0.659 2	0.011 1
StructDepth	NYU Depth V2	高视觉显著性	11	0.701 8	0.008 3
			15	0.665 1	0.008 3
			21	0.611 4	0.010 6
		低视觉显著性	11	0.689 6	0.008 0
			15	0.648 5	0.007 9
			21	0.592 8	0.010 2
	KITTI	高视觉显著性	11	0.706 7	0.008 2
			15	0.681 5	0.008 3
			21	0.651 4	0.009 7
		低视觉显著性	11	0.698 6	0.006 5
			15	0.672 9	0.007 0
			21	0.638 3	0.008 1

高显著性区域时, 测试框与其邻域的估计景深分布的相似性总体高于测试框位于低显著性区域的测试结果, 但随着测试框增大, 重叠指数均有降低趋势; 有监督景深估计模型 BTS 在不同区域类型中得到的重叠指数普遍高于自监督景深估计模型 StructDepth; NYU Depth V2 数据集上得到的重叠指数与 KITTI 数据集上的重叠指数非常相近; 不论将测试框置于高视觉显著性还是低视觉显著性区域, 测试框与其对应邻域的景深分布总具有较高的重叠度(重叠指数均大于 0.592 8).

第 3.1 节与第 3.2 节的测试结果表明, 在室内和室外 2 种场景下, 测试框和对应邻域内像素的估计景深分布均具有比较高的相似性, 并且该结论与所使用的视觉显著性估计模型和图像的场景类型等因素关联性不强.

3.3 测试框内像素及其邻域内像素景深估计的误差分析

本文使用 SalGAN 模型估计视觉显著性, 在 NYU Depth V2 和 KITTI 数据集上对 BTS 模型进行测试. 首先将测试框置于高(低)复杂度和高(低)显著性区域, 使用 BTS 模型估计出景深图; 然后根据景深真值计算对应的误差图; 最后在误差图上进行分析, 统计位于不同类型区域的测试框和其对应邻域内的均方根误差(root-mean-square error, RMSE), 并计算二者之间的重叠指数. 测试结果如表 4 和表 5 所示, 其中, RMSE 给出了均值和方差, 重叠指数仅给出了均值. 表 4 中, 由于 KITTI 数据集中的景深真值非常稀疏, 且要保证采样足够的像素进行重叠指数的计算, 因此将测试框边长设为 22×22 , 30×30 和 42×42 ; 邻域框边长为 31×31 ,

表 4 KITTI 数据集上图像高(低)复杂度和高(低)视觉显著性区域的景深估计误差对比

场景类型	测试框尺寸	测试框		邻域框		重叠指数
		均值	方差	均值	方差	
高复杂度	22	6.053 5	33.791 5	6.800 8	35.702 7	0.694 7
	30	7.524 9	38.481 8	9.260 1	44.027 4	0.676 2
	42	10.043 9	62.352 7	9.039 1	51.150 5	0.659 8
低复杂度	22	5.512 2	33.037 8	5.693 0	31.514 3	0.742 7
	30	5.725 8	30.030 1	5.902 5	30.116 4	0.726 7
	42	5.756 7	43.470 7	5.223 3	36.718 7	0.668 8
高视觉显著性	22	8.273 4	57.808 7	8.456 9	53.504 8	0.662 8
	30	9.698 4	65.242 9	10.275 6	59.317 5	0.652 7
	42	11.317 0	70.975 7	10.022 4	55.004 4	0.587 0
低视觉显著性	22	3.762 0	15.941 2	3.959 5	16.347 7	0.755 2
	30	4.235 9	19.116 0	4.280 4	18.674 3	0.739 2
	42	4.213 8	22.187 4	3.873 0	19.593 1	0.688 9

表 5 NYU Depth V2 数据集上图像高(低)复杂度和高(低)视觉显著性区域的景深估计误差对比

场景类型	测试框尺寸	测试框		邻域框		重叠指数
		均值	方差	均值	方差	
高复杂度	11	0.328 6	0.127 5	0.338 4	0.124 6	0.777 2
	15	0.326 9	0.095 5	0.336 1	0.094 1	0.759 9
	21	0.373 8	0.146 2	0.384 0	0.144 0	0.723 9
低复杂度	11	0.318 5	0.143 6	0.316 2	0.138 3	0.770 8
	15	0.322 3	0.131 8	0.317 3	0.125 8	0.743 5
	21	0.378 3	0.175 9	0.361 0	0.158 7	0.693 8
高视觉显著性	11	0.346 8	0.117 8	0.355 0	0.120 0	0.774 2
	15	0.350 3	0.117 5	0.356 7	0.117 1	0.755 0
	21	0.372 9	0.124 9	0.379 2	0.119 4	0.713 4
低视觉显著性	11	0.321 3	0.122 3	0.314 5	0.123 0	0.766 0
	15	0.330 8	0.128 5	0.328 5	0.127 5	0.737 7
	21	0.375 3	0.149 0	0.363 8	0.140 8	0.684 8

41×41 和 59×59; 采样数为 35, 60 和 90.

从表 4 和表 5 中可以看出, 不论测试框位于图像中何种类型的区域, 当测试框尺寸逐渐增大时, 误差的均值均略有增大, 误差分布的重叠指数也总是比较大(NYU DepthV2 数据集中大于 0.684 8, KITTI 数据集中大于 0.587 0); 在室内外场景下, 所得结论非常相似. 这些结果再次说明, 测试框内像素的估计景深, 与其邻域内像素的估计景深的分布具有比较明显的相似性.

4 图像视觉显著性与景深估计精度之间的关系分析

本文在 NYU Depth V2 和 KITTI 数据集上对 BTS 和 SC-DepthV3 这 2 种景深估计模型进行测试, 并使用 SalGAN 和 FastSal 模型估计测试图像的视觉显著性图, 对景深估计模型在不同显著性区域内的误差进行统计分析.

首先使用视觉显著性估计模型计算测试图像的显著性图, 并将显著值归一化后划分为 100 个等值区间; 然后计算景深估计误差图, 并通过掩模技术得到图像中不同显著性区间像素内的景深估计误差; 最后对数据进行统计分析, 确定景深估计模型在不同显著性区间内的性能.

理想情况下, 希望所使用的景深误差度量准则与物体景深无关, 因为只有“剔除”物体景深这一因素的影响, 才便于分析图像景深估计误差与图像视觉显著性之间的关系.

在针孔投影模型下, 摄像机坐标系下的空间点 (X, Y, Z) 与对应的图像平面点 (x, y) 的投影关系为

$$(x, y) = \left(\frac{X}{Z} f, \frac{Y}{Z} f \right).$$

其中, f 表示相机焦距. 假定图像在离散过程中, x 轴和 y 轴有相同的离散步长 Δs , 且像素坐标的 u 轴和 v 轴无扭曲(扭曲参数为 0), 则空间点 (X, Y, Z) 对应的像素坐标为

$$(u, v) = \left(\frac{x}{\Delta s}, \frac{y}{\Delta s} \right).$$

因此像素离散误差为

$$(\Delta u, \Delta v) \approx \left(-\frac{X}{Z^2 \Delta s} f, -\frac{Y}{Z^2 \Delta s} f \right) \Delta Z.$$

进一步, 假定 $\Delta u = \Delta v$, 则有

$$\Delta Z \approx \frac{\Delta s Z^2}{\sqrt{2} f} \sqrt{\frac{1}{X^2} + \frac{1}{Y^2}} \Delta u = \frac{\Delta s Z^2}{\sqrt{2} f X Y} \sqrt{X^2 + Y^2} \Delta u \quad (1)$$

假定像素 (u_i, v_i) 对应的图像坐标为 (x_i, y_i) , 则投影于图像坐标 (x_i, y_i) 的所有空间点 (X, Y, Z) 位于通过光心 $(0, 0, 0)$ 和 (x_i, y_i, f) 的直线上. 该直线的方程为

$$\begin{cases} \frac{Y}{y_i} = \frac{X}{x_i} \\ \frac{Z}{f} = \frac{Y}{y_i} \end{cases}.$$

将 $X = \frac{x_i}{f} Z$ 和 $Y = \frac{y_i}{f} Z$ 代入式(1), 可以得到

$$\Delta Z \approx \frac{\Delta s \sqrt{x_i^2 + y_i^2}}{\sqrt{2} x_i y_i} Z \Delta u \quad (2)$$

式(2)表明, 景深误差 ΔZ 正比于景深 Z , 因此理论上绝对相对误差(absolute relative error, Abs Rel)和对数均方根误差(root-mean-square error, RMSLE)是与景深 Z 无关的度量, 更合适度量景深估计误差与视觉显著性之间的关系, 其计算公式为

$$\text{Abs Rel} = \frac{1}{|N|} \sum_{i=1}^N \frac{\|d_i - d'_i\|}{d'_i},$$

$$\text{RMSLE} = \sqrt{\frac{1}{|N|} \sum_{i=1}^N \|\ln(d_i) - \ln(d'_i)\|^2}.$$

其中, d_i 为模型估计的景深; d' 为景深真值; N 为采样数量. 因此, 本文使用 Abs Rel, RMSLE 度量视觉显著性与景深估计误差之间的关系.

在不同视觉显著性估计模型上的测试结果相似, 图 2 所示为使用 SalGAN 模型估计视觉显著性图的结果, 其中, FastSal 模型下的测试结果见补充资料^①. 可以看出, 使用 Abs Rel 和 RMSLE 作为景深误差度量指标时, 在景深较低的室内场景中, 高视觉显著性区域的 RMSLE 均略低于低视觉显著性区域, 但 Abs Rel 的情况相反; 在室外场景中, 高视觉显著性区域的 2 种误差均略高于低视觉显著性区域; 总体来说, 这些由于模型和数据集导致的差异并不显著, 并且这 2 种误差的变化范围很小(Abs Rel 的变化幅度小于 0.13, RMSLE 的变化幅度小于 0.11). 因此可以认为, 模型的景深估计精度与图像显著性的相关性并不强.

① <https://github.com/DK05310214/What-Have-the-Single-Image-Based-Depth-Prediction-Models-Learnt>

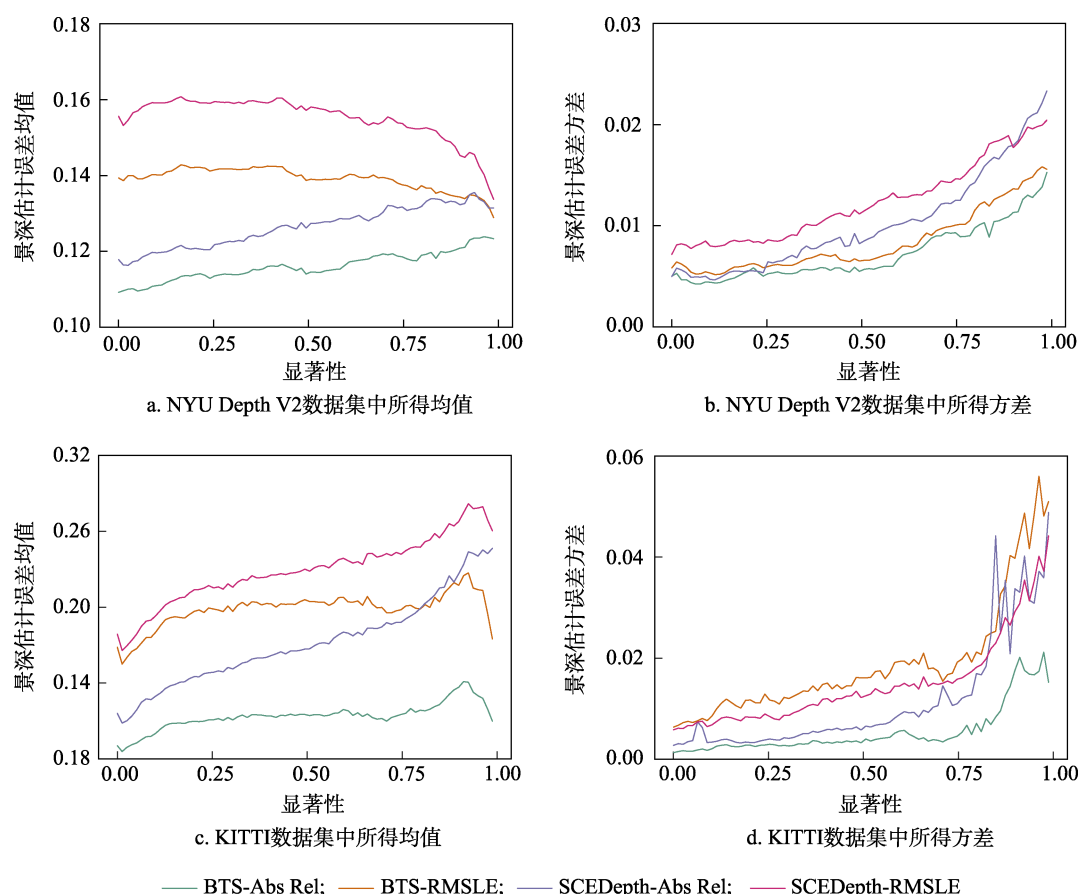


图 2 景深估计误差与图像视觉显著性之间的关系

需要说明的是, 对于将双目图像对作为训练数据的景深估计模型, 当景深比较小时, 训练图像对的 2 个相机之间的“公共视野”比较小, 如对 KITTI 数据集, 理论上, 当景深小于 0.27 m, 训练图像对的 2 个相机已没有“公共视野”, 此时“自监督学习框架”已失去其理论基础. 所以在景深比较小的区间, 模型的景深估计误差均比较大并且受图像离散效应的影响也比较大, 一般不能很好地体现模型的景深估计性能.

5 结论与讨论

本文从模型对图像无纹理区域估计的景深与其邻域的估计景深之间的关系, 以及图像中的视觉显著性与景深估计误差之间的关系这 2 个方面, 分析了基于深度学习的单目景深估计模型的可解性问题. 通过 3 种视觉显著性估计模型、3 种基于有监督的景深估计模型和 3 种自监督学习的景深估计模型, 在 3 个数据集上的测试结果表明, 当无纹理的测试框位于不同视觉显著性区域和不同图像复杂度区域时, 测试框内像素与其邻域内像素

的估计景深的分布都有比较高的相似性, 说明模型对这些无纹理、无视觉语义区域估计的景深, 很大程度上是其邻域区域的估计景深的某种填充效应; 另外, 使用 Abs Rel 和 RMSLE 作为景深误差度量指标时, 同一景深估计模型在不同数据集上的景深估计误差, 随图像视觉显著性的变化趋势确实存在一些差异, 不同的景深估计模型之间也确实存在一些差异. 但总体上, 这些由于模型和数据集导致的差异并不显著, 2 种误差的变化范围很小, 因此可以认为模型的景深估计精度与图像显著性的相关性并不强.

由于景深估计模型在无纹理区域依然可以估计景深, 因此无纹理区域的估计景深是其邻域估计景深的某种填充效应, 应该是可以预料到的结论, 本文的贡献在于定量地测试和分析了这个问题. 需要指出的是, 本文仅仅从 2 个侧面对单目景深估计模型的特性进行测试分析, 这些分析还比较初步. 例如, 本文度量了测试框区域与其邻域区域内像素估计景深的分布之间的重叠指数, 这种“分布度量”显然与像素在测试框中的位置无关, 从原理上而言, 任何填充效应均应该有空间效应,

即测试框内某个像素的估计景深应该受其较近邻域像素景深的影响大, 较远像素的影响小, 这种空间影响效应目前还未考虑, 这也是下一步拟开展的工作之一。

当前的测试结果表明, 单目景深估计模型的性能与图像的视觉显著性之间的关联性不强。由于景深估计通常是众多视觉任务的一个中间环节, 不论是场景三维重建还是场景分析, 人们总是对高视觉显著性区域更加感兴趣。因此今后在景深估计模型设计与训练时, 有必要充分地考虑图像视觉显著性效应, 如将视觉显著性图作为引导信息, 使模型更关注感兴趣的区域, 最终提高该区域的景深估计精度, 更好地服务于下游任务。

参考文献(References):

- [1] Masoumian A, Rashwan H A, Cristiano J, *et al.* Monocular depth estimation using deep learning: a review[J]. *Sensors*, 2022, 22(14): 5353
- [2] Khan F, Salahuddin S, Javidnia H. Deep learning-based monocular depth estimation methods-a state-of-the-art review[J]. *Sensors*, 2020, 20(8): 2272
- [3] Zhao C Q, Sun Q Y, Zhang C Z, *et al.* Monocular depth estimation based on deep learning: an overview[J]. *Science China Technological Sciences*, 2020, 63(9): 1612-1627
- [4] Ming Y, Meng X Y, Fan C X, *et al.* Deep learning for monocular depth estimation: a review[J]. *Neurocomputing*, 2021, 438: 14-33
- [5] Dong X S, Garratt M A, Anavatti S G, *et al.* Towards real-time monocular depth estimation for robotics: a survey[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(10): 16940-16961
- [6] Vyas P, Saxena C, Badapanda A, *et al.* Outdoor monocular depth estimation: a research review[OL]. [2022-12-29]. <https://arxiv.org/abs/2205.01399>
- [7] van Dijk T, De Croon G. How do neural networks see depth in single images?[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 2183-2191
- [8] Geiger A, Lenz P, Stiller C, *et al.* Vision meets robotics: the KITTI dataset[J]. *The International Journal of Robotics Research*, 2013, 32(11): 1231-1237
- [9] Hu J J, Zhang Y, Okatani T. Visualization of convolutional neural networks for monocular depth estimation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Los Alamitos: IEEE Computer Society Press, 2019: 3868-3877
- [10] Silberman N, Hoiem D, Kohli P, *et al.* Indoor segmentation and support inference from RGBD images[C] //Proceedings of the 12th European Conference on Computer Vision. Heidelberg: Springer, 2012: 746-760
- [11] Pan J T, Ferrer C C, McGuinness K, *et al.* SalGAN: visual saliency prediction with generative adversarial networks[OL]. [2022-12-29]. <https://arxiv.org/abs/1701.01081>
- [12] Hu F Y, McGuinness K. FastSal: a computationally efficient network for visual saliency prediction[C] //Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Los Alamitos: IEEE Computer Society Press, 2021: 9054-9061
- [13] Kroner A, Senden M, Driessens K, *et al.* Contextual encoder-decoder network for visual saliency prediction[J]. *Neural Networks*, 2020, 129: 261-270
- [14] Pastore M, Calcagni A. Measuring distribution similarities between samples: a distribution-free overlapping index[J]. *Frontiers in Psychology*, 2019, 10: 1089
- [15] Pastore M, Di Loro P A, Mingione M, *et al.* Overlapping: estimation of overlapping in empirical distributions[EB/OL]. [2022-12-29]. <https://cran.r-project.org/web/packages/overlapping/index.html>
- [16] Lee J H, Han M K, Ko D W, *et al.* From big to small: multi-scale local planar guidance for monocular depth estimation[OL]. [2022-12-29]. <https://arxiv.org/abs/1907.10326>
- [17] Yuan W H, Gu X D, Dai Z Z, *et al.* NeW CRFs: neural window fully-connected CRFs for monocular depth estimation[OL]. [2022-12-29]. <https://arxiv.org/abs/2203.01502>
- [18] Bhat F S, Alhashim I, Wonka P. AdaBins: depth estimation using adaptive bins[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 4008-4017
- [19] Li B Y, Huang Y, Liu Z Y, *et al.* StructDepth: leveraging the structural regularities for self-supervised indoor depth estimation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 12643-12653
- [20] Peng R, Wang R G, Lai Y W, *et al.* Excavating the potential capacity of self-supervised monocular depth estimation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 15540-15549
- [21] Sun L B, Bian J W, Zhan H Y, *et al.* SC-DepthV3: robust self-supervised monocular depth estimation for dynamic scenes[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(1): 497-508
- [22] Jiang M, Huang S S, Duan J Y, *et al.* SALICON: saliency in context[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1072-1080
- [23] Lin T Y, Maire M, Belongie S, *et al.* Microsoft COCO: common objects in context[C] //Proceedings of the 13th European Conference on Computer Vision. Heidelberg: Springer, 2014: 740-755