水下场景中的伪装计数

孙国磊 [1] 安照崇 [1] 刘云 [2] 刘策 [1]
Christos Sakaridis[1] 范登平 [1*] Luc Van Gool[1,3]
[1] CVL, ETH Zurich, [2] I2R, A*STAR, [3] VISICS, KU Leuven
guolei.sun@vision.ee.ethz.ch

Abstract

最近，在视觉社区中，难以分辨的场景理解引起了很多关注。本文通过系统研究一个名为难以分辨物体计数（*IOC*）的新挑战，进一步推进了这一领域的前沿。其目标是计算与周围环境相似的物体数量。由于缺乏适当的 *IOC* 数据集，本文提出了一个大规模数据集 ***IOCfish5K***，其中包含共 *5,637* 个高分辨率图像和 *659,024* 个注释的中心点。本文提出的数据集包含大量水下场景中难以分辨的物体（主要是鱼类），使得注释过程更具挑战性。*IOCfish5K* 比现有的难以分辨场景数据集更具优势，因为它具有更大的规模、更高的图像分辨率、更多的注释和更密集的场景。所有这些方面使其成为迄今为止最具挑战性的 *IOC* 数据集，支持这一领域的发展。为了进行基准测试，本文选择了 *14* 种主流的物体计数方法，并在 *IOCfish5K* 上进行了仔细评估。此外，本文提出了 ***IOCFormer***，一个新的强大基线模型，它在统一的框架中结合了密度和回归分支，并能够有效地处理隐蔽场景下的物体计数。实验证明，*IOCFormer* 在 *IOCfish5K* 上取得了最先进的性能。相关资源可在*github.com/GuoleiSun/Indiscernible-Object-Counting*上找到。

1. Introduction

对象计数——估计图像中对象实例的数量——一直是计算机视觉中的一个重要主题。理解场景中每个类别的数量对于智能代理在其环境中导航来说至关重要。这个任务可以是最终目标，也可以是辅助步骤。对于后者来说，计数对象已被证明对于实例分割 [14]、动作定位 [54] 和行人检测 [84] 有所帮助。对于前者来说，它是监控 [79]、人群监测 [6]、野生动物保护 [56]、饮食模式理解 [55] 和细胞种群分析 [1] 中的核心算法。

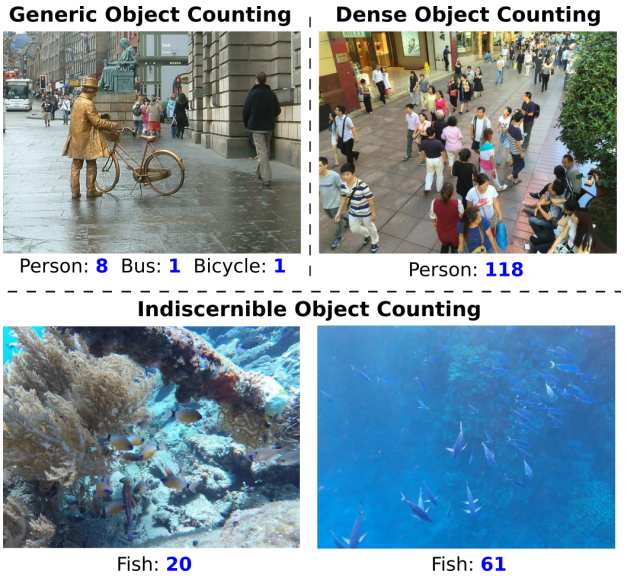

图 1. 不同计数任务的示意图。左上方：通用物体计数（GOC），用于在自然场景中计数各种类别的物体。右上方：密集物体计数（DOC），用于在密集场景中计数前景类别的物体。下方：不可分辨物体计数（IOC），用于在不可分辨场景中计数前景类别的物体。你能在给定的示例中找到所有的鱼吗？对于 GOC、DOC 和 IOC，所显示的图像分别来自 PASCAL VOC [18]、ShanghaiTech [92] 和新的 IOCfish5K 数据集。

之前的目标计数研究主要遵循两个方向：通用/常见物体计数（GOC）[8,14,32,68] 和密集物体计数（DOC）[28, 36, 50, 57, 64, 67, 92]。这两个子任务的区别在于所

*通讯作者：范登平 (dengpfan@gmail.com)，此论文为 CVPR2023 论文 [69] 翻译件.

数据集	年份	伪装场景	图片数量	平均分辨率	自由视野	计数统计				网站
						总和	最小	平均	最大	
UCSD [6]	2008	✗	2,000	158×238	✗	49,885	11	25	46	Link
Mall [10]	2012	✗	2,000	480×640	✗	62,325	13	31	53	Link
UCF_CC_50 [27]	2013	✗	50	2101×2888	✓	63,974	94	1,279	4,543	Link
WorldExpo'10 [91]	2016	✗	3,980	576×720	✗	199,923	1	50	253	Link
ShanghaiTech B [92]	2016	✗	716	768×1024	✗	88,488	9	123	578	Link
ShanghaiTech A [92]	2016	✗	482	589×868	✓	241,677	33	501	3,139	Link
UCF-QNRF [28]	2018	✗	1,535	2013×2902	✓	1,251,642	49	815	12,865	Link
Crowd_surv [88]	2019	✗	13,945	840×1342	✗	386,513	2	35	1420	Link
GCC (synthetic) [81]	2019	✗	15,212	1080×1920	✗	7,625,843	0	501	3,995	Link
JHU-CROWD++ [65]	2019	✗	4,372	910×1430	✓	1,515,005	0	346	25,791	Link
NWPU-Crowd [80]	2020	✗	5,109	2191×3209	✓	2,133,375	0	418	20,033	Link
NC4K [51]	2021	✓	4,121	530×709	✓	4,584	1	1	8	Link
CAMO++ [33]	2021	✓	5,500	N/A	✓	32,756	N/A	6	N/A	Link
COD [19]	2022	✓	5,066	737×964	✓	5,899	1	1	8	Link
IOCfish5K (Ours)	2023	✓	5,637	1080×1920	✓	659,024	0	117	2,371	Link

表 1. 现有数据集中用于密集目标计数（DOC）和不可区分目标计数（IOC）的统计数据。

研究的场景，如图 1 所示。GOC 解决的是在自然/常见场景 [8]（即来自 PASCAL VOC [18] 和 COCO [41] 的图像）中计数各类物体的问题。需要估计的物体数量通常较少，即少于 10 个。另一方面，DOC 主要计数拥挤场景中前景类别的物体。估计的计数可以是几百甚至几万。被计数的对象通常是人（人群计数）[36, 39, 89]、车辆 [26, 57] 或植物 [50]。得益于大规模数据集 [10, 18, 28, 65, 80, 92] 和在这些数据集上训练的深度卷积神经网络（CNNs），在 GOC 和 DOC 方面取得了重要进展。然而，据我们所知，目前没有关于计数难以区分的物体的相关工作。

在不可分辨的场景下，前景物体与背景具有相似的外观、颜色或纹理，因此很难被传统视觉系统检测到。这种现象存在于自然和人工场景中 [20, 33]。因此，对不可分辨场景的场景理解自从一些开创性的工作出现以来就受到了越来越多的关注 [20, 34]。研究者已经提出和形式化了各种任务：伪装对象检测（COD）[20]，伪装实例分割（CIS）[33] 和伪装视频对象检测（VCOD）[12, 31]。然而，还没有先前的研究关注在不可分辨的场景中对象计数，这是一个重要的方面。

本文研究了新的不可辨识物体计数（**IOC**）任务，该任务侧重于计算不可辨识场景中的前景物体数量。图 1展示了这一挑战。诸如图像分类 [17, 24]、语义分割 [11, 42] 和实例分割 [3, 23] 等任务的进展都归功于大规模数据集的可用性 [16, 18, 41]。同样，一个高质量的 IOC 数据集将促进其发展。尽管现有的数据集 [20, 33, 51] 中包含了实例级别的注释，可用于 IOC，但它们存在以下限制：1）这些数据集中注释的对象总数有限，图像分辨率较低；2）它们仅包含有少实例数的场景/图像；3）实例级别的掩码注释可以通过计算质心转换为点监督，但计算出的点不一定落在对象内部。

为了促进关于 IOC 的研究，本文构建了一个大规模数据集，IOCfish5K。本文收集了 5,637 张具有不可辨识场景的图像，并标注出了 659,024 个中心点。与现有数据集相比，提出的 IOCfish5K 具有以下几个优势：1）它是 IOC 领域中最大规模的数据集，包括图像数量、图像分辨率和总物体数；2）IOCfish5K 中的图像经过精心筛选，包含多样的不可辨识场景；3）点注释准确，位于每个物体的中心位置。我们将我们的数据集与现有的 DOC 和 IOC 数据集进行了比较，如表 1所示，并在图 2中展示了示例图像。

基于提出的 IOCfish5K 数据集，本文对 14 个主流基准方法 [32, 36, 39, 40, 45, 47, 52, 66, 74, 77, 90, 92] 进行了系统研究。本文发现在现有的 DOC 数据集上表现良好的方法不一定在我们具有挑战性的数据集上保持竞争力。因此，本文提出了一种名为 IOCFormer 的简单而有效的方法。具体而言，本文结合了基于密度的方法 [77] 和基于回归的方法 [39] 的优点。前者可以估计图像中的目标密度，而后者直接回归点的坐标，简单而优雅。IOCFormer 包含两个分支：密度和回归。来自密度分支的密度感知特征通过提出的密度增强 Transformer 编码器（DETE）帮助突出不可区分的对象。然后，经

过传统的 Transformer 解码器处理后，生成预测的目标点。实验表明，IOCFormer 优于所有考虑的算法，证明了其在 IOC 上的有效性。

总结一下，本文的贡献有三个方面。

- 本文提出了新的不可辨对象计数（IOC）任务。为了促进对 IOC 的研究，本文提供了一个大规模数据集 IOCfish5K ，包含 5,637 张图像和 659,024 个准确的点标签。
- 本文选择了 14 种经典和高性能的对象计数方法，并在提出的 IOCfish5K 上对它们进行评估，以便进行基准测试。
- 本文提出了一种新颖的基准模型，即 IOCFormer ，它将基于密度和回归的方法集成到一个统一的框架中。此外，本文还提出了一种新颖的基于密度的 Transformer 编码器，逐渐利用密度分支中的密度信息来帮助检测不可辨对象。

2. Related Works

2.1. 通用物体计数

通用/常见物体计数（GOC）[14]，也被称为日常物体计数 [8]，是指在自然场景中计算各种类别物体实例的数量。GOC 的流行基准是 PASCAL VOC [18] 和 COCO [41]。该任务首次在开创性工作 [8] 中提出并研究，该工作将图像分割为不重叠的块，并通过估算其计数来预测物体实例的数量。LC [14] 使用图像级别计数监督为每个类别生成密度图，提高了计数性能和实例分割。RLC [15] 通过仅要求对训练类别的子集提供计数信息而不是所有类别，进一步减少了监督。与之不同的是，LCFCN [32] 利用点级监督，并输出每个物体实例的单个斑点。

2.2. 稠密物体计数

稠密物体计数（DOC）[13,28,50,53,57,63,64,83,85,86,92] 用于计算稠密场景中物体的数量。DOC 包括众多任务，如人群计数 [28,29,37,44,64,78,80,87,92,94]、车辆计数 [26,57]、植物计数 [50]、细胞计数 [1] 和企鹅计数 [2]。其中，人群计数，即计算人的数量，引起了最大的关注。人群计数的流行基准包括 ShanghaiTech [92]、UCF-QNRF [28]、JHU-CROWD++ [64]、NWPU-Crowd [80] 和 Mall [10]。对于车辆计数，研究人员主要使用 TRANCOS [57]、PUCPR+ [26] 和 CAPRK [26]。对于其他类别的 DOC，可用的数据集包括用于植物计数的 MTC [50]，用于细胞计数的 CBC [1]，以及用于企鹅计数的 Penguins [2]。DOC 与 GOC 不同之处在于 DOC 要计数的对象更多，主要关注一类特定的物体。

以计数策略为基础，先前的 DOC（密度人群计数）工作可分为三组：检测方法 [21,35,43,61]，回归方法 [6,7,27,39,66] 和密度图生成方法 [36,40,47,49,62,70,77]。检测方法首先检测对象，然后进行计数。虽然直观，但在拥挤场景中的检测表现较差。回归方法通过将全局特征回归到整个图像的计数 [6,7,27] 或直接将局部特征回归到点坐标 [39,66] 来进行计数。大部分先前的研究都集中在学习密度图上，密度图是具有降低空间尺寸的单通道输出。它表示每个位置上对象的分数数量，并且空间整合后等于图像中对象的总数。密度图可以通过使用高斯核生成的伪密度图 [36,45,73] 或直接使用标注点图 [52,70,77] 进行学习。

对于架构选择，过去在密集人群计数（DOC）方面的工作也可以分为基于 CNN 的方法 [32,36,47,48,60,66] 和基于 Transformer 的方法 [38,39,70]。从本质上讲，卷积神经网络（CNNs）具有有限的感受野，并且仅使用局部信息。相比之下，Transformer 可以建立特征之间的长程/全局关系。Transformer 在 DOC 方面的优势已被 [38,59,70] 证明。

2.3. 无法分辨的目标计数

最近，无法分辨的场景理解变得很受欢迎 [19,31,33,33,34,93]。它包含一组任务，专注于检测、实例分割和视频对象检测/分割。其目标是分析具有难以在视觉上识别的对象的场景 [20,31]。

本文研究了无法分辨的目标计数（IOC）这一新任务，它处于密集目标计数（DOC）和无法分辨的场景理解的交叉点上。最近提出的数据集 [19,33,51]，用于隐藏场景理解，可以通过将实例级别的掩码转换为点来用作 IOC 的基准。然而，它们有一些限制，如 §1中所讨论的。因此，本文提出了 IOC 的首个大规模数据集，即 IOCfish5K。

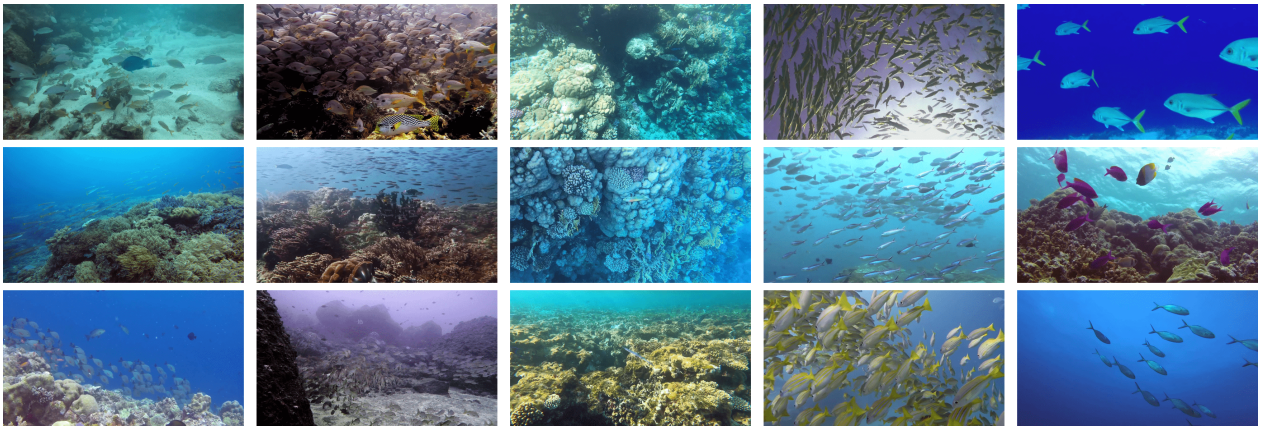

图 2. 来自 IOCfish5K 的示例图像。从左列到右列：典型样本，难以区分且密集的样本，难以区分且不那么密集的样本，不太难以区分且密集的样本，不太难以区分且不那么密集的样本。

3. IOCfish5K 数据集

3.1. 图片收集

水下场景包含许多难以辨认的物体（海马、暗石鱼、狮子鱼和藻海龙），这是因为能见度有限和积极的拟态。因此，本文专注于收集水下场景的图像。

本文首先收集了 YouTube 上的水下场景视频，使用了一般关键词（水下场景、海洋潜水、深海场景等）和特定类别的关键词（乌贼、拟态章鱼、灯笼鱼、石鱼等）。总共，本文收集了 135 个高质量的视频，长度从几十秒到几个小时不等。然后，我们每隔 100 帧（3.3 秒）保留一张图像，以避免重复。这仍然产生了大量的图像，其中一些显示相似的场景或质量较低。因此，在图像收集的最后一步，6 位专业标注员仔细审查了数据集，并删除了那些不优质的图像。最终的数据集包含 5,637 张图像，其中一些图像如图 2所示。这个步骤总共耗时 200 个人工小时。

3.2. 图片标注

标注原则。 目标是在每个动物的可见部分中心注释一个点。本文追求准确性和完整性。前者表示注释点应该放置在对象中心，并且每个点对应一个对象实例。后者意味着没有任何对象应该被漏掉注释。

标注工具。 为了简化标注过程，本文开发了一个基于开源 Labelimg 工具[1]的工具。它提供以下功能：通过点击在图像中生成点注释、拖动/删除点、遇到困难情况时标记点，以及缩放功能。这些功能帮助标注人员产生高质量的点注释，并通过讨论标记的案例来解决不确定性。

[1] https://github.com/heartexlabs/labelImg

Datasets	# IMG (0-50)	# IMG (51-100)	# IMG (101-200)	# IMG (>200)	Total
NC4K [51]	4,121	0	0	0	4,121
COD [19]	5,066	0	0	0	5,066
IOCfish5K	2,663	1,000	957	1,017	5,637

表 2. 关于数据集在不同密度（计数）范围内图像分布的比较。本文计算了每个数据集在四个密度范围内的图像数量。

注释过程。 整个过程分为三个步骤。首先，所有标注员（6 名专家）接受培训，熟悉自己的任务。他们接受关于海洋动物和经过良好标注的样本的指导。然后，每个标注员被要求对 50 张图像进行注释。注释被检查和评估。当一个标注员通过评估后，他/她可以进入下一步。第二，图像被分配给 6 个标注员，每个标注员负责数据集的一部分。标注员们被要求讨论不确定的情况，并达成共识。最后，他们在两轮中检查和完善注释。第二步耗时 600 个人工小时，而第三步中的每一轮检查耗时 300 个小时。注释过程的总成本为 1,200 个人工小时。

3.3. 数据集细节

提议的 IOCfish5K 数据集包含 5,637 张高质量图像，标注有 659,024 个点。表 2显示了每个计数范围内的图像数量（0-50，51-100，101-200 和 200 以上）。在 IOCfish5K 的所有图像中，有 957 张具有中等到高密度的对象，即 101 到 200 个实例之间。此外，1,017 张图像（数据集的 18%）显示出非常密集的场景（每张图

像超过 200 个对象）。

为了在 IOCfish5K 上进行标准化的基准测试，本文将其随机分为三个不重叠的部分：训练集（3,137）、验证集（500）和测试集（2,000）。对于每个划分，图像在不同计数范围内的分布都遵循相似的分布。

表 1比较了 IOCfish5K 与先前数据集的统计数据。IOCfish5K 相对于现有数据集的优势有四个。**(1)** IOCfish5K 是面向不可分辨场景的最大规模对象计数数据集。在大小、图像分辨率和注释点数量方面，它优于其对应的数据集，如 NC4K [51]、CAMO++ [33] 和 COD [19]。例如，最大的现有 IOC 数据集 CAMO++ [33] 包含总共 32,756 个对象，而 IOCfish5K 中有 659,024 个点。**(2)** IOCfish5K 具有更密集的图像，这使其成为目前最具挑战性的 IOC 基准。如表 2所示，有 1,974 张图像中有超过 100 个对象。**(3)** 尽管 IOCfish5K 专门用于 IOC，但它比现有 DOC 数据集也具有一些优势。例如，与 JHU-CROWD++ [64]（其中之一最大规模的 DOC 基准）相比，本文的数据集包含更多分辨率更高的图像。**(4)** IOCfish5K 专注于具有海洋动物注释的水下场景，这使其与表 1中显示的所有现有数据集不同。因此，该数据集对于 DOC 的迁移学习和领域自适应也具有价值 [9, 22, 25, 46]。

4. IOCFormer

我们首先介绍我们提出的 IOCFormer 模型的网络结构，它由密度分支和回归分支组成。然后，我们解释了一种新颖的密度增强 Transformer 编码器，该编码器旨在帮助网络更好地识别和检测难以辨别的对象。

4.1. 网络结构

正如前面提到的，目前主流的目标计数方法可以分为两组：基于密度计数方法 [36, 77] 和基于回归计数方法 [39, 66]。基于密度的方法 [36, 77] 通过学习图像中估计的目标密度来建立一个密度图。而基于回归的方法 [39, 66] 则直接回归目标中心点的坐标，更加直观和简洁。至于 IOC，前景物体由于其颜色和纹理上的相似性，很难与背景区分开。基于密度的方法能够估计目标密度水平，这个特性可以用来突出（难以区分的）前景物体并提高基于回归的方法的性能。换句话说，基于密度和基于回归的方法的优点可以结合起来。因此，本文提出了 IOCFormer，其中包含两个分支：密度分支和回归分支，如图 3 所示。密度分支的信息有助于优化回归分支的特征。

给定了一个输入图像 $\boldsymbol{I}$，其中包含真实的对象点 $\{(x_i, y_i)\}_{i=1}^{K}$，其中 (x_i, y_i) 表示第 i 个对象点的坐标，K 是对象的总数。目标是训练一个对象计数模型，预测图像中对象的数量。本文首先通过将图像输入编码器来提取特征图 $\boldsymbol{F} \in \mathbb{R}^{h \times w \times c_1}$（其中 h、w 和 c_1 分别表示高度、宽度和通道数），接下来，$\boldsymbol{F}$ 经过密度和回归分支的处理。

密度分支将 $\boldsymbol{F}$ 输入到由两个 3×3 卷积核组成的卷积解码器中。获得密度感知特征图 $\boldsymbol{F}_d \in \mathbb{R}^{h \times w \times c_2}$，其中 c_2 是通道数。然后，密度头（一个具有 1×1 卷积核和 ReLU 激活的卷积层）将 $\boldsymbol{F}_d$ 映射到一个单通道密度图 $\boldsymbol{D} \in \mathbb{R}^{h \times w}$，其值为非负数。与 [77] 类似，密度分支中使用的计数损失（L_1 损失）定义为：

$$\mathcal{L}_D = |\|\boldsymbol{D}\|_1 - K|, \tag{1}$$

其中 $\|\cdot\|_1$ 表示矩阵的逐元素 L_1 范数。密度图 $\boldsymbol{D}$ 估计了空间维度上的物体密度水平。因此，密度头之前的特征图 $\boldsymbol{F}_d$ 是密度感知的，并包含物体密度信息，可以利用这些信息加强具有不可辨别物体实例的特征区域。

就回归分支而言，来自编码器的特征图 $\boldsymbol{F}$ 和来自密度分支的密度感知特征图 $\boldsymbol{F}_d$ 首先被送到我们的密度增强 Transformer 编码器中，详细描述见 §4.2。在这个模块之后，经过改进的特征与目标查询一起传递到典型的 Transformer 解码器中 [72]。解码的查询嵌入被分类头和回归头用于生成预测。详细信息在 §4.3 中解释。

4.2. 密度增强 Transformer 编码器

这里我们详细解释密度增强 Transformer 编码器 (DETE)。典型 Transformer 编码器 (TTE) 和提出的 DETE 的结构如图 4所示。DETE 与 TTE 不同，它处理两个输入：由初始编码器提取的特征 ($\boldsymbol{F}$) 和来自密度分支的密度感知特征 ($\boldsymbol{F}_d$)。DETE 使用密度感知特征图来改进编码器特征图。通过获取图像中密集分布对象和稀疏分布对象的区域信息，回归分支可以更准确地预测无法区分的对象实例的位置。

我们首先将 $\boldsymbol{F}$ 投影到 $\hat{\boldsymbol{F}} \in \mathbb{R}^{h \times w \times c}$，并将 $\boldsymbol{F}_d$ 投影到 $\hat{\boldsymbol{F}}_d \in \mathbb{R}^{h \times w \times c}$，使用一个 MLP 层，使得通道数量 ($c$) 匹配。第一个 Transformer 层的输入是 $\hat{\boldsymbol{F}}$，$\hat{\boldsymbol{F}}_d$ 和

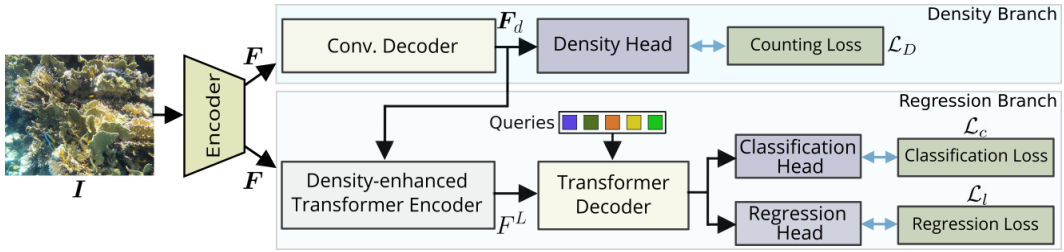

图 3. IOCFormer 的概述。给定输入图像，本文使用编码器提取特征图，该特征图由密度分支和回归分支进行处理。密度增强的 Transformer 编码器利用来自密度分支的物体密度信息生成更相关的回归特征。更多详细信息请参阅 §4。

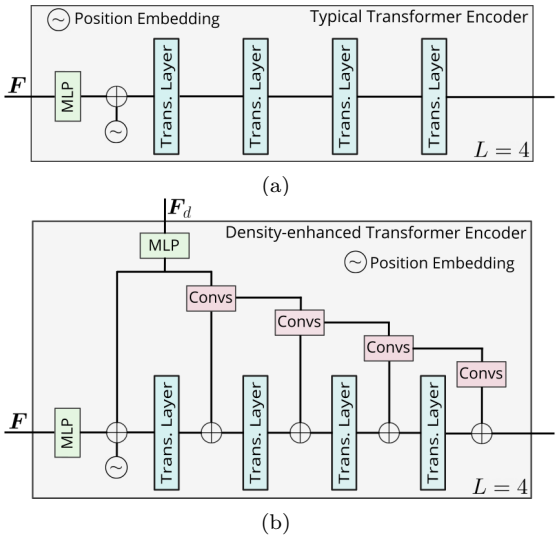

图 4. 当 $L = 4$ 时，典型的 Transformer 编码器 (a) 和本文的密度增强 Transformer 编码器 (b) 的比较。

位置嵌入 $E \in \mathbb{R}^{hw \times c}$ 的组合。这个过程可以表示为：

$$F^1 = \mathrm{Rs}(\hat{\boldsymbol{F}}) + \mathrm{Rs}(\hat{\boldsymbol{F}}_d) + E; \quad F^2 = \mathrm{Trans}(F^1), \tag{2}$$

其中，$\mathrm{Rs}(\cdot)$ 表示将特征图展平其空间维度的操作，$\mathrm{Trans}(\cdot)$ 表示一个 Transformer 层。在此之后，使用额外的 Transformer 层进一步优化特征，如下所示：

$$\begin{aligned} \boldsymbol{F}_d^1 &= \hat{\boldsymbol{F}}_d, \\ \boldsymbol{F}_d^i &= \mathrm{Convs}(\boldsymbol{F}_d^{i-1}), \quad i = 2, 3, ..., L-1, \\ F^{i+1} &= \mathrm{Trans}(F^i + \mathrm{Rs}(\boldsymbol{F}_d^i)), \quad i = 2, 3, ..., L-1, \end{aligned} \tag{3}$$

其中 $\mathrm{Convs}(\cdot)$ 表示包含两个卷积层的卷积块。Transformer 层的总数为 L，这也表示将 Transformer 和卷积特征合并的总次数。在 Equ. (3) 之后，我们获得密度优化的特征 $F^L \in \mathbb{R}^{hw \times c}$，将其传递给 Transformer 解码器。

我们的 DETE 的好处也可以从*全局*和*局部*信息的角度来解释。在 Equ. (3) 中的每个 Transformer 层之前，我们会将前一个 Transformer 层的特征（全局）和卷积块的特征（局部）进行合并。在这个过程中，全局和局部信息逐渐融合，从而提升了模型的表示能力。

4.3. 损失函数

在 DETE 模块之后，我们获得密度精化特征 F^L。接下来，Transformer 解码器将精化特征 F^L 和可训练的查询嵌入 $Q \in \mathbb{R}^{n \times c}$（包含 n 个查询）作为输入，并输出解码嵌入 $\hat{Q} \in \mathbb{R}^{n \times c}$。Transformer 解码器包含多个层，每个层都包含一个自注意力模块、一个交叉注意力层和一个前馈神经网络（FFN）。有关更多详细信息，请参阅经典工作 [72]。$\hat{Q}$ 包含 n 个解码表示，对应于 n 个查询。根据 [39]，每个查询嵌入通过分类头和回归头被映射到一个置信分数和一个点坐标。设 $\{p_i, (\hat{x}_i, \hat{y}_i)\}_{i=1}^n$ 表示所有查询的预测，其中 p_i 是确定点属于前景的预测置信分数，$(\hat{x}_i, \hat{y}_i)$ 是第 i 个查询的预测坐标。然后，我们对预测 $\{p_i, (\hat{x}_i, \hat{y}_i)\}_{i=1}^n$ 和真值 $\{(x_i, y_i)\}_{i=1}^K$ 进行匈牙利匹配 [4,39]。注意，n 大于 K，以便每个真值点都有一个匹配的预测。匈牙利匹配基于 k 最近邻匹配目标 [39]。具体而言，匹配成本取决于预测点和真值点之间的距离、预测点的置信分数以及预测点和真值点平均最近邻距离的差异 [39]。匹配后，我们计算分类损失 $\mathcal{L}_c$，增强匹配的预测的置信分数，并抑制未匹配的预测的置信分数。为了监督预测的坐标学习，我们还计算定位损失 $\mathcal{L}_l$，衡量匹配的预测坐标与相应的真值坐标之间的 L_1 距离。有关更多详细信息，请参阅 [39]。最终损失函数定义为：

$$\mathcal{L} = \lambda \mathcal{L}_D + \mathcal{L}_c + \mathcal{L}_l, \tag{4}$$

其中 λ 被设置为 0.5。密度和回归分支是使用 Equ. (4) 共同训练的。在推断过程中，我们使用回归分支的预测结果。

方法	出版	验证集 (500)			测试集 (2,000)		
		MAE↓	MSE↓	NAE↓	MAE↓	MSE↓	NAE↓
MCNN [92]	CVPR'16	81.62	152.09	3.53	72.93	129.43	4.90
CSRNet [36]	CVPR'18	43.05	78.46	1.91	38.12	69.75	2.48
LCFCN [32]	ECCV'18	31.99	81.12	0.77	28.05	68.24	1.12
CAN [47]	CVPR'19	47.77	83.67	2.10	42.02	74.46	2.58
DSSI-Net [45]	ICCV'19	33.77	80.08	1.25	31.04	69.11	1.68
BL [52]	ICCV'19	19.67	44.21	0.39	20.03	46.08	0.55
NoisyCC [74]	NeurIPS'20	19.48	41.76	0.39	19.73	46.85	0.46
DM-Count [77]	NeurIPS'20	19.65	42.56	0.42	19.52	45.52	0.55
GL [75]	CVPR'21	18.13	44.57	0.33	18.80	46.19	0.47
P2PNet [66]	ICCV'21	21.38	45.12	0.39	20.74	47.90	0.48
KDMG [76]	TPAMI'22	22.79	47.32	0.90	22.79	49.94	1.17
MPS [90]	ICASSP'22	34.68	59.46	2.06	33.55	55.02	2.61
MAN [40]	CVPR'22	24.36	40.65	2.39	25.82	45.82	3.16
CLTR [39]	ECCV'22	17.47	37.06	0.29	18.07	41.90	0.43
IOCFormer (Ours)	CVPR'23	**15.91**	**34.08**	**0.26**	**17.12**	**41.25**	**0.38**

表 3. 在验证集和测试集上与最先进方法进行比较。最佳结果以**粗体**突出显示。

5. 实验

5.1. 实验设置

比较的模型。 由于没有专门为 IOC 设计的算法，我们选择了 14 种最近的开源 DOC 方法进行基准测试。选定的方法包括：MCNN [92]、CSRNet [36]、LCFCN [32]、CAN [47]、DSSI-Net [45]、BL [52]、NoisyCC [74]、DM-Count [77]、GL [75]、P2PNet [66]、KDMG [76]、MPS [90]、MAN [40] 和 CLTR [39]。其中，P2PNet 和 CLTR 基于回归，而其他方法基于密度图估计。"

实施细节。 对于 MCNN 和 CAN 等方法，本文在实验中使用了开源的重新实现版本。对于其他方法，本文使用官方代码和默认参数。所有实验都在 PyTorch [58] 和 NVIDIA GPU 上进行。在 DETE 中，L 设置为 6，查询的数量 (n) 设置为 700。根据 [39]，本文的 IOCFormer 使用 ResNet-50 [24] 作为编码器，预训练于 Imagenet [16]。其他模块/参数是随机初始化的。对于数据增强，本文使用随机调整大小和水平翻转。图像被随机裁剪为 256×256 的输入。每个批次包含 8 张图片，并使用 Adam 优化器 [30]。在推理过程中，本文将图像分割成与训练过程中相同大小的块。根据 [39]，本文使用一个阈值 (0.35) 来过滤掉背景预测。

指标。 为了评估基线方法和提出的方法的有效性，本文根据 [39, 77, 80] 计算所有图像的预测计数与真实计数之间的平均绝对误差 (MAE)，平均平方误差 (MSE) 和平均归一化绝对误差 (NAE)。

5.2. 计数结果与分析

我们在表格 3 中展示了 14 个主流的人群计数算法和 IOCFormer 的结果。所有方法都遵循相同的评估协议：通过验证集选择模型。根据结果，我们观察到：

- 在所有先前的方法中，最近的 CLTR [39] 在 MAE、MSE 和 NAE 的测试集上分别达到 18.07、41.90 和 0.43，优于其他方法。原因是该方法使用 Transformer 编码器学习全局信息，并使用 Transformer 解码器直接预测目标实例的中心点。
- 一些方法 (MAN [40] 和 P2PNet [66]) 在 JHU++ [65] 和 NWPU [80] 等 DOC 数据集上表现出竞争力，但在 IOCfish5K 上表现较差。举例来说，MAN 在 JHU++ 上的 MAE 和 MSE 分别达到 53.4 和 209.9，优于其他方法，包括 CLTR 在 MAE 和 MSE 上分别达到 59.5 和 240.6。然而，与 CLTR、DM-Count、NoisyCC 和 BL 相比，MAN 在 IOCfish5K 上表现不佳。这表明为 DOC 设计的方法未必适用于不可分辨的对象。因此，IOC 需要特定设计的解决方案。
- 这些方法，包括 BL、NoisyCC、DM-Count 和 GL，提出了适用于人群计数的新损失函数，尽管简单但表现良好。例如，GL 在测试集上的 MAE、MSE 和 NAE 分别达到 18.80、46.19 和 0.47。

不同于先前的方法，IOCFormer 专门针对 IOC 进行了两个创新设计：(1) 在统一框架中结合了密度和回归分支，改进了底层特征；(2) 基于密度的 Transformer

Methods	DETE	MAE↓	MSE↓	NAE↓
DB	✗	18.25	39.77	0.29
Regression	✗	17.47	37.06	0.29
DB+Regression	✗	16.94	35.92	**0.26**
	✓	**15.91**	**34.08**	**0.26**

表 4. 密度分支（DB）和 DETE 对 IOCfish5K 验证集性能的影响。对于没有使用 DETE 的 DB+Regression，使用了一个典型的 Transformer 编码器（TTE）。

L	MAE↓	MSE↓	NAE↓
2	16.75	35.87	0.28
4	16.59	35.23	0.26
6	15.91	34.08	0.26
8	**15.72**	**33.63**	**0.24**

表 5. DETE 中的 Transformer 层数或卷积块数量的影响

编码器，增强了物体存在的特征区域。在验证集和测试集上，IOCFormer 在 MAE、MSE 和 NAE 方面均优于所有先前的方法。除了定量结果，我们还在图 5中展示了一些方法的定性结果。

5.3. 消融实验

密度分支和 DETE 的影响。 正如前面提到的，所提出的模型将密度分支和回归分支结合在一个统一的框架中，旨在充分利用它们的优势。在表 4中，我们展示了分别训练密度分支和回归分支的结果。我们还提供了同时训练密度分支和回归分支但不使用提出的 DETE 的结果。比较表明，回归分支，虽然简单直接，但表现优于仅使用密度分支。

此外，训练两个分支但不使用 DETE 的结果比仅使用回归分支的性能更好。这种改进可以从多任务学习的角度解释 [5, 71, 82]。额外的密度分支可以被视为一项“附加任务”，有助于编码器学习更好的特征。通过建立密度分支和回归分支之间的连接，可以获得更好的性能。与不使用 DETE 的变体相比，我们的最终模型将平均绝对误差 (MAE) 从 16.94 减少到 15.91，均方误差 (MSE) 从 35.92 减少到 34.08，具有明显的优势。结果验证了 DETE 通过利用密度分支生成的信息来增强特征的有效性。

L 的影响。 我们改变 DETE 中 Trans 或 Convs 的数量，并在表 5中报告结果。通过增加 L，我们获得更好的性能，表明我们的 DETE 能够产生相关特征。在我们的主要设置中，我们使用 $L = 6$ 来平衡复杂性和性能。

6. 结论与未来工作

本文对一项名为不可区分目标计数（IOC）的新挑战进行了严格的研究，该挑战专注于在不可区分场景中预测物体数量。为了解决缺乏大规模数据集的问题，本文提供了高质量的 IOCfish5K 数据集，主要包含水下场景，并在对象（主要是鱼类）实例的中心位置进行了点注释。本文选择了多个现有的主流基准模型，并在 IOCfish5K 上进行了评估，证明了密集物体计数（DOC）和 IOC 之间存在领域差距。此外，本文提出了一种专门针对 IOC 的方法，名为 IOCFormer，它具有两个新设计：将密度和回归分支结合在一个统一的模型中，以及一种密度增强的 Transformer 编码器，将对象密度信息从密度分支传递到回归分支。IOCFormer 在 IOCfish5K 上取得了最好的性能。总之，本文的数据集和方法为未来的研究人员提供了一个深入探索这一新任务的机会。

未来工作。有几个方向。(1) 提高性能和效率。虽然本文的方法达到了最先进的性能，但在 IOCfish5K 上进一步改进 MAE、MSE 和 NAE 方面的计数结果还有改进的空间。此外，在实际应用中部署计数模型时，效率也很重要。(2) 研究 IOC 和 DOC 之间的领域自适应。DOC 数据集，相对于 IOC 数据集更多，如何利用可用的 DOC 数据集来改进 IOC 是一个实际问题。(3) 获得一个通用的计数模型，可以计数所有的东西（人、植物、细胞、鱼等）。

致谢

Christos Sakaridis 和 Deng-Ping Fan 受到 Toyota Motor Europe（TRACE-Zürich 研究项目）的资助。

参考文献

[1] Mohammad Mahmudul Alam and Mohammad Tariqul Islam. Machine learning approach of automatic identification and counting of blood cells. *HTL*, 6(4):103–108, 2019. 1, 3

[2] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *ECCV*, 2016. 3

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *IEEE ICCV*, 2019. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey

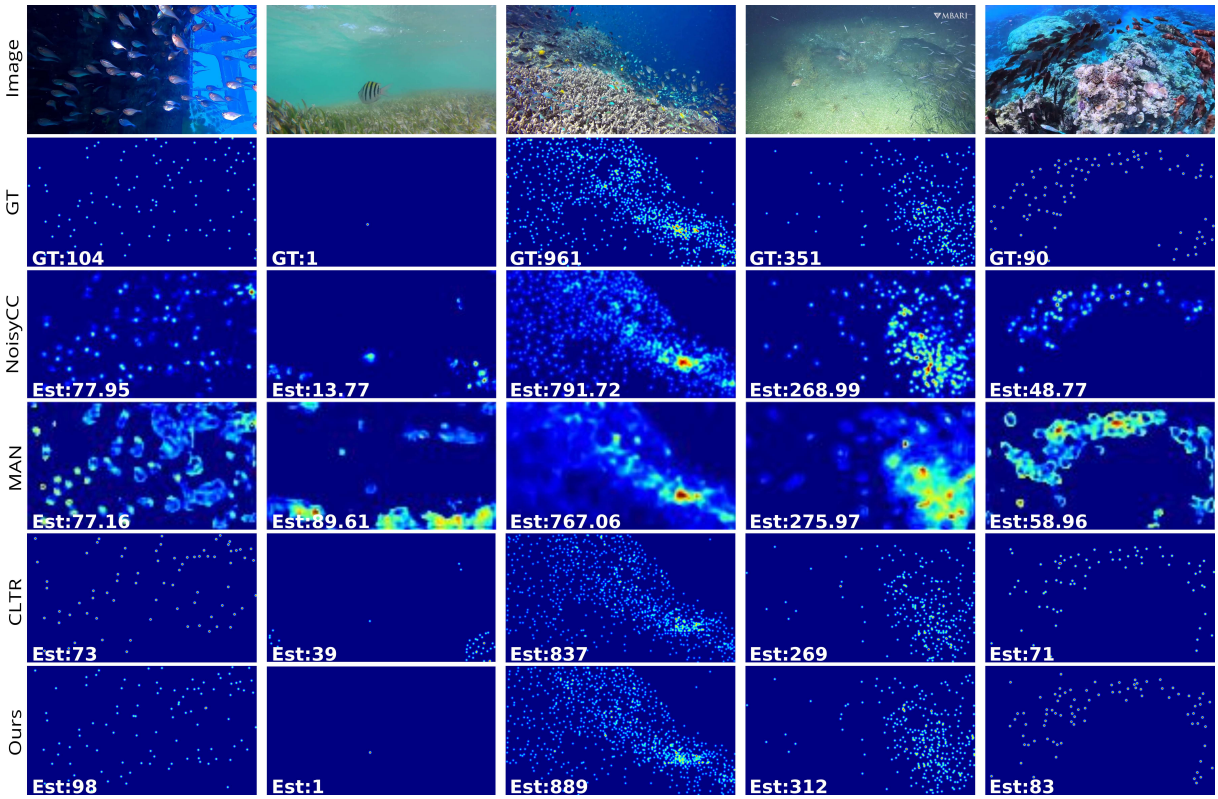

图 5. 定性比较各种算法（NoisyCC [74]，MAN [40]，CLTR [39] 和我们的）。每种情况的真实或估计计数显示在左下角。最佳观看效果请进行缩放。

Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 6

[5] Rich Caruana. Multitask learning. *ML*, 28(1):41–75, 1997. 8

[6] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE CVPR*, 2008. 1, 2, 3

[7] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *IEEE ICCV*, 2009. 3

[8] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *IEEE CVPR*, 2017. 1, 2, 3

[9] Binghui Chen, Zhaoyi Yan, Ke Li, Pengyu Li, Biao Wang, Wangmeng Zuo, and Lei Zhang. Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting. In *IEEE ICCV*, 2021. 5

[10] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012. 2, 3

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017. 2

[12] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *IEEE CVPR*, 2022. 2

[13] Zhi-Qi Cheng, Qi Dai, Hong Li, Jingkuan Song, Xiao Wu, and Alexander G. Hauptmann. Rethinking spatial invariance of convolutional networks for object counting. In *IEEE CVPR*, 2022. 3

[14] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *IEEE CVPR*, 2019. 1, 3

[15] Hisham Cholakkal, Guolei Sun, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Luc Van Gool. Towards partial supervision for generic object counting in natural scenes. *IEEE TPAMI*, 44(3):1604–1622, 2022. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, 2009. 2, 7

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[18] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 1, 2, 3

[19] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10):6024–6042, 2022. 2, 3, 4, 5

[20] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE CVPR*, 2020. 2, 3

[21] Weina Ge and Robert T Collins. Marked point processes for crowd counting. In *IEEE CVPR*, 2009. 3

[22] Shenjian Gong, Shanshan Zhang, Jian Yang, Dengxin Dai, and Bernt Schiele. Bi-level alignment for cross-domain crowd counting. In *IEEE CVPR*, 2022. 5

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE ICCV*, 2017. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016. 2, 7

[25] Yuhang He, Zhiheng Ma, Xing Wei, Xiaopeng Hong, Wei Ke, and Yihong Gong. Error-aware density isomorphism reconstruction for unsupervised cross-domain crowd counting. In *AAAI*, 2021. 5

[26] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *IEEE ICCV*, 2017. 2, 3

[27] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *IEEE CVPR*, 2013. 2, 3

[28] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, 2018. 1, 2, 3

[29] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *IEEE CVPR*, 2020. 3

[30] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7

[31] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *ACCV*, 2020. 2, 3

[32] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *ECCV*, 2018. 1, 2, 3, 7

[33] Trung-Nghia Le, Yubo Cao, Tan-Cong Nguyen, Minh-Quan Le, Khanh-Duy Nguyen, Thanh-Toan Do, Minh-Triet Tran, and Tam V Nguyen. Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. *IEEE TIP*, 31:287–300, 2021. 2, 3, 5

[34] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 184:45–56, 2019. 2, 3

[35] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *IEEE ICPR*, 2008. 3

[36] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE CVPR*, 2018. 1, 2, 3, 5, 7

[37] Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, and Shenghua Gao. Locating and counting heads in crowds with a depth prior. *IEEE TPAMI*, 44(12):9056–9072, 2022. 3

[38] Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. Transcrowd: weakly-supervised crowd counting with transformers. *SCIS*, 65(6):1–14, 2022. 3

[39] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *ECCV*, 2022. 2, 3, 5, 6, 7, 9

[40] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multi-faceted attention. In *IEEE CVPR*, 2022. 2, 3, 7, 9

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 3

[42] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *IEEE CVPR*, 2019. 2

[43] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *IEEE CVPR*, 2018. 3

[44] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, and Chunhua Shen. Weighing counts: Sequential crowd counting by reinforcement learning. In *ECCV*, 2020. 3

[45] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *IEEE ICCV*, 2019. 2, 3, 7

[46] Weizhe Liu, Nikita Durasov, and Pascal Fua. Leveraging self-supervision for cross-domain crowd counting. In *IEEE CVPR*, 2022. 5

[47] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *IEEE CVPR*, 2019. 2, 3, 7

[48] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Exploiting unlabeled data in CNNs by self-supervised learning to rank. *IEEE TPAMI*, 41(8):1862–1878, 2019. 3

[49] Xiyang Liu, Jie Yang, Wenrui Ding, Tieqiang Wang, Zhijin Wang, and Junjun Xiong. Adaptive mixture regression network with local counting map for crowd counting. In *ECCV*, 2020. 3

[50] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. Tasselnet: counting maize tassels in the wild via local counts regression network. *PM*, 13(1):79, 2017. 1, 2, 3

[51] Yunqiu Lyu, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *IEEE CVPR*, 2021. 2, 3, 4, 5

[52] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *IEEE ICCV*, 2019. 2, 3, 7

[53] Chenlin Meng, Enci Liu, Willie Neiswanger, Jiaming Song, Marshall Burke, David Lobell, and Stefano Ermon. Is-count: Large-scale object counting from satellite images with covariate-based importance sampling. *arXiv preprint arXiv:2112.09126*, 2021. 3

[54] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *IEEE ICCV*, 2019. 1

[55] Huu-Thanh Nguyen, Chong-Wah Ngo, and Wing-Kwong Chan. Sibnet: Food instance counting and segmentation. *PR*, 124:108470, 2022. 1

[56] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *PNAS*, 115(25):E5716–E5725, 2018. 1

[57] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, 2016. 1, 2, 3

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 7

[59] Yifei Qian, Liangfei Zhang, Xiaopeng Hong, Carl R Donovan, and Ognjen Arandjelovic. Segmentation assisted u-shaped multi-scale transformer for crowd counting. In *BMVC*, 2022. 3

[60] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *ECCV*, 2018. 3

[61] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R Venkatesh Babu. Locate, size, and count: accurately resolving people in dense crowds via detection. *IEEE TPAMI*, 43(8):2739–2751, 2020. 3

[62] Zenglin Shi, Pascal Mettes, and Cees GM Snoek. Counting with focus for free. In *IEEE ICCV*, 2019. 3

[63] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *IEEE CVPR*, 2022. 3

[64] Vishwanath Sindagi, Rajeev Yasarla, and Vishal MM Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE TPAMI*, 44(5):2594–2609, 2022. 1, 3, 5

[65] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *IEEE ICCV*, 2019. 2, 7

[66] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *IEEE ICCV*, 2021. 2, 3, 5, 7

[67] Qingyu Song, Changan Wang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Jian Wu, and Jiayi Ma. To choose or to fuse? scale selection for crowd counting. In *AAAI*, 2021. 1

[68] Tobias Stahl, Silvia L Pintea, and Jan C Van Gemert. Divide and count: Generic object counting by image divisions. *IEEE TIP*, 28(2):1035–1044, 2018. 1

[69] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *IEEE CVPR*, 2023. 1

[70] Guolei Sun, Yun Liu, Thomas Probst, Danda Pani Paudel, Nikola Popovic, and Luc Van Gool. Boosting crowd counting with transformers. *arXiv preprint arXiv:2105.10926*, 2021. 3

[71] Guolei Sun, Thomas Probst, Danda Pani Paudel, Nikola Popović, Menelaos Kanakis, Jagruti Patel, Dengxin Dai, and Luc Van Gool. Task switching network for multi-task learning. In *IEEE ICCV*, 2021. 8

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5, 6

[73] Jia Wan and Antoni Chan. Adaptive density map generation for crowd counting. In *IEEE ICCV*, 2019. 3

[74] Jia Wan and Antoni Chan. Modeling noisy annotations for crowd counting. In *NeurIPS*, 2020. 2, 7, 9

[75] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *IEEE CVPR*, 2021. 7

[76] Jia Wan, Qingzhong Wang, and Antoni B. Chan. Kernel-based density map generation for dense object counting. *IEEE TPAMI*, 44(3):1357–1370, 2022. 7

[77] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. In *NeurIPS*, 2020. 2, 3, 5, 7

[78] Changan Wang, Qingyu Song, Boshen Zhang, Yabiao Wang, Ying Tai, Xuyi Hu, Chengjie Wang, Jilin Li, Jiayi Ma, and Yang Wu. Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. In *IEEE ICCV*, 2021. 3

[79] Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *IEEE CVPR*, 2011. 1

[80] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE TPAMI*, 2020. 2, 3, 7

[81] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *IEEE CVPR*, 2019. 2

[82] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *ICML*, 2009. 8

[83] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *IEEE CVPR*, 2021. 3

[84] Jin Xie, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Mubarak Shah. Count-and similarity-aware r-cnn for pedestrian detection. In *ECCV*, 2020. 1

[85] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *IEEE ICCV*, 2019. 3

[86] Haipeng Xiong and Angela Yao. Discrete-constrained regression for local counting models. In *ECCV*, 2022. 3

[87] Chenfeng Xu, Dingkang Liang, Yongchao Xu, Song Bai, Wei Zhan, Xiang Bai, and Masayoshi Tomizuka.

Autoscale: Learning to scale for crowd counting. *IJCV*, 130(2):405–434, 2022. 3

[88] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. In *IEEE ICCV*, 2019. 2

[89] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *ECCV*, 2020. 2

[90] Mohsen Zand, Haleh Damirchi, Andrew Farley, Mahdiyar Molahasani, Michael Greenspan, and Ali Etemad. Multiscale crowd counting and localization by multitask point supervision. In *IEEE ICASSP*, 2022. 2, 7

[91] Cong Zhang, Kai Kang, Hongsheng Li, Xiaogang Wang, Rong Xie, and Xiaokang Yang. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE TMM*, 18(6):1048–1061, 2016. 2

[92] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE CVPR*, 2016. 1, 2, 3, 7

[93] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *IEEE CVPR*, 2022. 3

[94] Joey Tianyi Zhou, Le Zhang, Jiawei Du, Xi Peng, Zhiwen Fang, Zhe Xiao, and Hongyuan Zhu. Locality-aware crowd counting. *IEEE TPAMI*, 44(7):3602–3613, 2022. 3