高精度图像二类分割

秦雪彬[1], 戴航[1], 胡晓彬[2], 范登平 *[3], 邵岭[4], Luc Van Gool[3]

[1] 穆罕默德 · 本 · 扎耶德人工智能大学, 阿布扎比, 阿拉伯联合酋长国
[2] 腾讯优图实验室, 上海, 中国
[3] 苏黎世联邦理工学院, 苏黎世, 瑞士
[4] 特斯联科技集团, 中国
[5] xuebin@ualberta.ca, hang.dai@mbzuai.ac.ae, xiaobin.hu@tum.de, dengpfan@gmail.com, ling.shao@ieee.org, vangool@vision.ee.ethz.ch

摘要 本文系统地研究了一项新任务，称为图像二类分割 (DIS)，旨在从自然图像中分割出高精度的物体。为此，本文收集了第一个大规模 DIS 数据集，称为 DIS5K，其中包含 5470 张高分辨率 (如 2K、4K 或更大) 图像，覆盖了各种背景中伪装的、显著的或结构稠密的物体。DIS 使用非常精细化的标签进行注释。此外，本文引入了一个简单的中间监督基线 (IS-Net)，使用特征级和掩码级指导来进行 DIS 模型训练。IS-Net 在 DIS5K 上的表现优于各种前沿基线，这使其成为一个通用的自学习监督网络，可以促进 DIS 的未来研究。此外，本文设计了一种称为人工矫正量 (HCE) 的新指标，它近似于校正假阳性和假阴性所需的鼠标点击操作次数。HCE 被用来衡量模型和现实应用之间的差距，因此是现有指标的补充。最后，本文进行了最大规模的基准测试，评估了 16 个代表性的分割模型，提供了关于对象复杂性的深入讨论，并展示了几个潜在的应用 (例如背景去除、艺术设计和 3D 重建)。希望这些努力可以为学术界和工业界开辟有前景的发展方向。项目主页: https://xuebinqin.github.io/dis/index.html。

Keywords: 图像二类分割, 高分辨率, 衡量标准

1 引言

目前，驱动大量人工智能 (AI)模型的计算机视觉数据集的标注精度在一定程度上满足了机器感知系统的要求。然而，人工智能已经步入了一个要求计算机视觉算法输出高度精确从而支持精细的人机交互时代。相比于分类 [15, 39, 74] 和检测 [29, 30, 69]，分割可以为广泛的应用提供更精确的几何目标描述，如图像编辑 [31]、AR/VR [65]、医学图像分析 [71] 和机器人操作 [7] 等。

* 通信作者: 范登平。本文为 ECCV2022 [64] 的中文翻译版。由俞珍妮翻译，范登平、胡晓彬校稿。

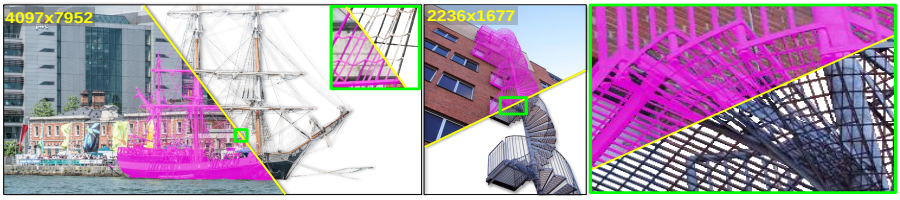


图 1: DIS5K 数据集的样本图像。放大以获得最佳视觉效果。

这些应用程序可以根据它们对现实世界物体的直接影响，分为“轻”(如图像编辑和分析) 和“重”(如人机交互)。“轻”的 (图1)，通常可以后期修正，允许相对宽容的分割错误。而在“重”的分割中，分割偏差或失败更有可能对物体造成物理损伤或伤害 (有时是致命的) 人类，因此，需要高精度和鲁棒性好的模型。目前，大多数的分割模型由于准确性和鲁棒性问题，在那些“重”应用中仍然不太适用。因此，本文的目标是在一个被称为图像二类分割 (DIS) 的通用框架中解决“重”和“轻”的应用问题，它旨在分割高度精确的物体。

现有的分割任务主要关注具有特定特征的对象，如显著的 [80, 83, 95]、伪装的 [23, 40, 75]、结构稠密的 [45, 91] 或特定种类的 [38, 46, 55, 71, 73]。它们具有相同的输入/输出格式，并且排它性机制很少用于在它们的模型中分割特定目标，这意味着它们通常是依赖于数据集的。因此，本文提议在无二义性的标注下定义一个类别无关的 DIS 任务，以便准确地分割具有不同结构复杂性的对象，无论它们的特征如何。与语义分割 [14, 17, 47, 63, 104] 相比，本文的 DIS 任务主要针对单个或少数目标的图像，因此从这些目标中提取出更丰富的细节信息是可行的。因此，本文总结了四个**贡献**:

(i) 一个大规模、可扩展的 DIS 数据集 DIS5K，包含 5,470 幅高分辨率图像，并配有高度精确的二值分割掩码。

(ii) 一种新的基于中间监督的基线 IS-Net，它通过强制高维特征的直接同步来减少过拟合。

(iii) 一种新设计的人工矫正量 (HCE) 指标，通过计算修正错误区域所需的人类干预程度来衡量模型预测和现实应用之间的差距。

(iv) 在新的 DIS5K 的基础上，本文建立了完整的 DIS 基准，使本文的 DIS 调研最广泛。本文将 IS-Net 与 16 个前沿的分割模型进行了比较，展示了 IS-Net 良好的性能。

2 相关工作

深度学习时代的图像分割**任务与数据集**密切相关。一些分割任务，比如 [12,21,45,46,55,73,83,91]，甚至直接建立在数据集上。这些任务的定义完全相同，即: $P = F(\theta, I)$, 其中 I 和 P 分别表示输入图像和二值化输出图像。然而，研究这些任务之间的相关性却很少，这限制了其训练好的模型应用到更广的领域。此外，不同任务所使用的数据集也不具有排他性，这说明将图像二类分割 (DIS) 任务统一是可能的。**模型**常常在更强的表达能力和更高的过拟合风险之间挣扎。为了获得更具代表性的特征，研究人员设计了基于全卷积网络模型 [49]、编码-解码模型 [3,71]、粗到细模型 [84]、预测-优化模型 [67,80] 和视觉变换器模型 [102]。此外，为了平衡性能和时间成本，许多实时模型 [24,37,43,58,59,93,98] 被设计出来。其他方法，如权值正则化 [32]、dropout 法 [76]、密集监督 [41,66,88] 和混合损失 [50,67,100]，专注于缓解过拟合。密集监督是减少过拟合最有效的方法之一。然而，监控中间深度特征的侧输出可能不是最好的选择，因为从多通道深度特征到单通道侧输出的转换削弱了监督效果。**评价指标**可以被分为基于区域的 (如 IoU 与 Jaccard 指数 [1]、F-measure [13,70] 或者 Dice 系数 [78]、加权 F-measure [52])、基于边界的 (比如 CM [57]、边界 F-measure [16, 53, 56, 62, 67, 72, 97]、边界 IoU [9]、边界位移误差 (BDE) [27] 和 Hausdorff 距离 [4,5,34])、基于结构的 (比如 S-measure [19] 和 E-measure [20,22])、基于置信度的 (比如 MAE [61]) 等。它们主要是从数学或认知的角度来衡量预测结果和真值之间的一致性。但是，在实际应用中，将预测结果与满足需求的结果之间的同步成本指标还没有得到充分的研究。

3 本文的 DIS5K 数据集

3.1 数据收集与标注

数据收集。为了解决数据问题 (见章节2)，本文构建了一个高度精确的 DIS 数据集，名为 DIS5K。本文首先根据预先设计的关键字[6]从 Flickr[7]手动收集了 12000 多张图片。然后，根据物体结构的复杂性，共获得 5470 张图片分为 22 组、225 类 (图2) 。注意，选择策略类似于 Zhou 等人 [103] 的方法。大多数选定的图像只包含单个目标，以获得丰富而高精度的结构和细

[6] 由于这项研究的长期目标是促进机器与我们的生活/工作环境之间的“安全”和“高效”互动，这些关键词大多与我们日常生活中的普通物品 (比如自行车、椅子、背包、电缆和树等) 相关。

[7] 图片获得“允许商业使用和编辑”的许可

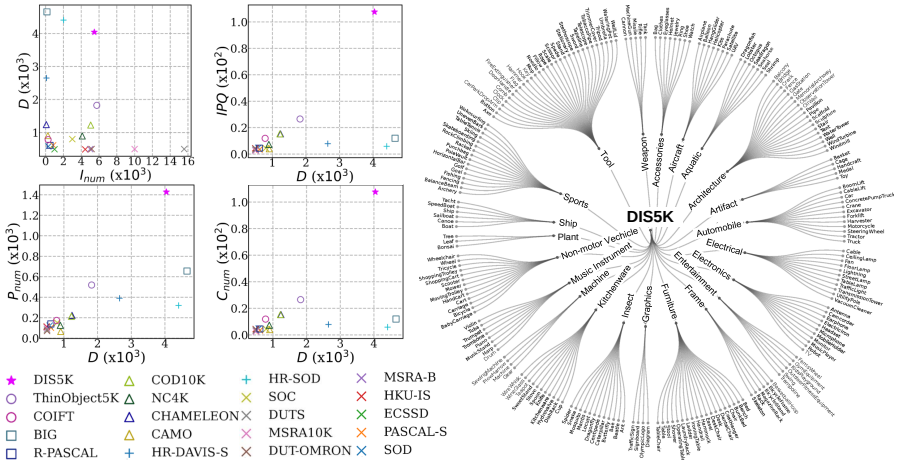


图 2: **左**: 不同复杂性之间的相关性。**右**: DIS5K 数据集的类别和组别。放大以获得最佳视觉效果。详见章节3.1.

节。同时，最大限度地避免了不同类别的多个对象同时出现造成的分割和标注混乱。具体而言，图像选择标准可归纳如下:

- 覆盖更多的类别，同时减少那些已经包含在现有其他数据集中具有简单结构的“冗余”样本的数量。
- 通过增加更多样化类内图像 (图3-f) 来扩大所选类别的类内差异 (见补充材料 (SM)中的 2.3 节)。
- 囊括更多结构复杂的类别, 比如栅栏、楼梯、电缆、盆景和树等，这些在我们的生活中很常见，但没有很好地被标记 (图3-a) 或由于标记困难而被其他数据集忽略。

因此，本文在 DIS5K 中所标注的目标主要是“预先设计的关键词所定义的图像前景对象”而不考虑它们的特征，比如显著、普通、伪装和结构稠密等。**数据标注**。使用 GIMP[8]对 DIS5K 的每张图像进行像素级精度手工标记。平均每张图像的标记时间约 30 分钟，有些图像的标记时间长达 10 小时。值得注意的是，有些标注的真值 (GT) 掩膜在视觉上接近图像抠图真值。标注对象有透明的和半透明的，并采用高达单个像素的二进制掩码进行标记。在这里，DIS 任务是不知道类别的，而本文的 DIS5K 是根据预先设计的关键字/类别收集的，这似乎是矛盾的。原因有三。(1) 关键词大大方便了大规

[8] https://www.gimp.org/

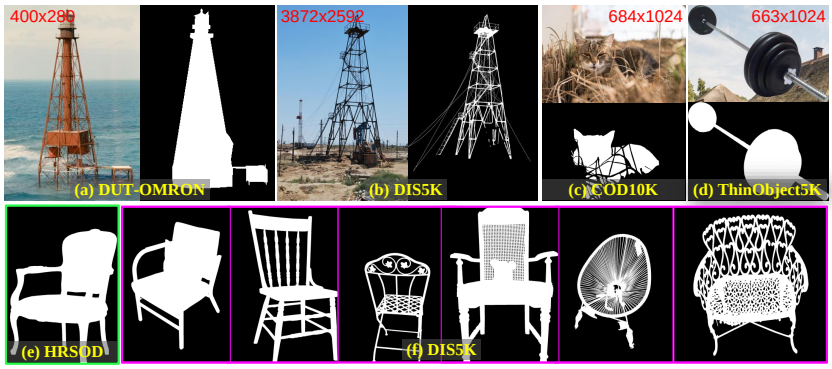


图 3: 不同数据集的定性比较。(a) 和 (b) 表明本文的 DIS5K 提供了更精确的标签。(c) 为 COD10K [23] 的一个样本，其结构复杂性是由遮挡引起的。(d) 展示了合成的 ThinObject5K [45] 数据集。(e) 和 (f) 表明 DIS5K 类内结构复杂性的多样性更大。

模数据集的检索和组织。(2) 为了达到类别不可知分割的目的，需要多样化的样本。根据类别来收集样本是保证数据集多样性下限的合理方法。本文的 DIS5K 的多样性上界是由大量样本的多样性特征 (比如纹理、结构、形状、对比和复杂性等) 决定的，保证了类别不可知分割的鲁棒性和泛化性。(3) 没有完美的数据集，因此对现有数据集进行重新组织或进一步扩展对于不同的实际应用程序来说通常是必要的。类别信息将极大地促进对已收集和待收集样本的追踪。因此，基于类别的数据收集与 DIS 任务的目标并不矛盾，而是内在一致的。

3.2 数据分析

为了更深入地了解 DIS 数据集，本文将 DIS5K 与其他 19 个相关数据集进行比较，包括: 9 组显著目标检测 (SOD) 数据集: SOD [57]、PASCAL-S [44]、ECSSD [90]、HKU-IS [42]、MSRA-B [48]、DUT-OMRON [92]、MSRA10K [12]、DUTS [83] 和 SOC [18]; (2) 两个高分辨率显著性目标检测 (HR-SOD) 数据集: HR-SOD [95] 和 HR-DAVIS-S [62, 95]; (3)4 个伪装目标检测 (COD) 数据集: CAMO [40]、CHAMELEON [75]、COD10K [23] 和 NC4K [51]; 两个语义分割 (SMS)[9]数据集: R-PASCAL [11, 17] 和 BIG [11]; 两个纤细目标分割 (TOS) 数据集: COIFT [45] 和 ThinObject5K [45]。主要从图像数量、图像维数和物体复杂度三个方面进行比较，如表 1 所示。

[9] 值得注意的是，这里只包括 R-PASCAL 和 BIG 数据集，因为它们的目标是高度精确的分割，并且它们的大多数图像包含一个或两个对象，这与列出的任务和数据集具有可比性。

表 1: 现有数据集的数据分析。详见章节3.2。

任务	数据集	数量	图像维度			对象复杂度		
		I_{num}	$H \pm \sigma_H$	$W \pm \sigma_W$	$D \pm \sigma_D$	$IPQ \pm \sigma_{IPQ}$	$C_{num} \pm \sigma_C$	$P_{num} \pm \sigma_P$
SOD	SOD [57]	300	366.87 ± 72.35	435.13 ± 72.35	578.28 ± 0.00	4.74 ± 3.89	2.25 ± 1.76	122.79 ± 62.97
	PASCAL-S [44]	850	387.63 ± 64.65	467.82 ± 61.46	613.22 ± 32.00	3.39 ± 2.46	5.14 ± 11.72	102.76 ± 70.09
	ECSSD [90]	1000	311.11 ± 56.27	375.45 ± 47.70	492.75 ± 19.78	3.26 ± 2.62	1.69 ± 1.42	107.54 ± 53.09
	HKU-IS [42]	4447	292.42 ± 51.13	386.64 ± 37.42	488.00 ± 29.44	4.41 ± 4.28	2.21 ± 2.07	114.05 ± 55.06
	MSRA-B [48]	5000	321.94 ± 56.33	370.86 ± 50.84	496.42 ± 22.53	2.89 ± 3.67	1.77 ± 2.25	102.04 ± 56.50
	DUT-OMRON [92]	5168	320.93 ± 54.35	376.78 ± 46.02	499.50 ± 22.97	4.08 ± 6.20	2.27 ± 3.54	71.09 ± 59.60
	MSRA10K [12]	10000	324.51 ± 56.26	370.27 ± 50.25	497.57 ± 22.79	2.54 ± 2.62	4.07 ± 17.94	101.95 ± 63.24
	DUTS [83]	15572	322.1 ± 53.69	375.48 ± 47.03	499.35 ± 21.95	3.37 ± 4.28	2.62 ± 4.73	84.78 ± 57.74
	SOC [18]	3000	480.00 ± 0.00	640.00 ± 0.00	800.00 ± 0.00	4.44 ± 3.57	13.69 ± 30.41	151.72 ± 154.83
HRS	HR-SOD [95]	2010	2713.12 ± 1041.7	3411.81 ± 1407.56	4405.40 ± 1631.03	5.85 ± 12.60	6.33 ± 16.65	319.32 ± 264.20
	HR-DAVIS-S [62]	92	1299.13 ± 440.77	2309.57 ± 783.59	2649.87 ± 899.05	7.84 ± 5.69	15.60 ± 29.51	389.58 ± 309.29
COD	CAMO [40]	250	564.22 ± 402.12	693.89 ± 578.53	905.51 ± 690.12	3.97 ± 4.47	1.48 ± 1.18	65.21 ± 40.99
	CHAMELEON [75]	76	741.80 ± 452.25	981.08 ± 464.88	1239.98 ± 629.19	15.25 ± 51.43	10.28 ± 48.03	222.45 ± 332.22
	NC4K [23]	4121	529.61 ± 158.16	709.19 ± 198.90	893.23 ± 223.94	7.28 ± 11.28	4.32 ± 9.44	125.43 ± 123.76
	COD10K [23]	5066	737.37 ± 185.65	963.85 ± 222.73	1224.53 ± 239.40	15.28 ± 71.84	17.18 ± 183.87	214.12 ± 857.83
SMS	R-PASCAL [11]	501	384.34 ± 64.69	469.66 ± 60.04	612.19 ± 36.32	4.44 ± 6.91	7.30 ± 8.73	139.31 ± 104.60
	BIG [11]	150	2801.11 ± 889.78	3672.43 ± 1128.90	4655.81 ± 1312.44	11.94 ± 31.43	31.69 ± 71.94	655.68 ± 710.20
TOS	COIFT [45]	280	488.27 ± 92.25	600.40 ± 78.66	782.73 ± 30.45	11.88 ± 12.5	4.01 ± 3.98	173.14 ± 74.54
	ThinObject5K [45]	5748	1185.59 ± 909.53	1325.06 ± 958.43	1823.03 ± 1258.49	26.53 ± 119.98	33.06 ± 216.07	519.14 ± 1298.54
DIS	**DIS5K (Ours)**	5470	2513.37 ± 1053.40	3111.44± 1359.51	4041.93 ± 1618.26	107.60 ± 320.69	106.84 ± 436.88	1427.82 ± 3326.72

图像维度对于分割任务至关重要，因为它对分割的准确性、效率和计算成本都有重要影响。图像高度、宽度、对角线长度的均值 (H, W, D) 和标准差 (σ_H, σ_W, σ_D) 见表1。BIG 数据集拥有最大的平均图像维度，但它只包含 150 幅图像。HR-SOD 的维度略大于本文的数据集，但复杂度较低。DIS5K 的平均维数几乎是 SOD 和 COD 数据集的 8 倍。此外，COD 数据集中的目标主要是动物和昆虫，这限制了其在多样化任务中的应用。

对象复杂度由等周不等式商 (IPQ) [60, 85, 91]、物体轮廓的数量 (C_{num}) 和主导点的数量 P_{num} 三个度量来描述。IPQ 主要将整体结构复杂度描述为 $IPQ = \frac{L^2}{4\pi A}$，其中 L 和 A 分别表示物体周长和区域面积。它被设计用来区分细长构件/薄凹结构的物体和近凸的物体。C_{num} 用来表示等高线层次上的拓扑复杂性，用于观测由许多 (小) 等高线组成的物体，这些物体通常对 IPQ 的影响较小。为了在更精细的层次上描述对象的复杂性，本文使用 P_{num} 来计算沿对象边界的主导点的数量 [68]。因此，沿边界的小锯齿段的复杂性通常不能用 IPQ 和 C_{num} 精确测量，而 P_{num} 可以很好地评价。本质上，P_{num} 是逼近分割掩码所需的多边形角的总数，它也直接反映了人工标记成本。因此，它被用于本文的人工矫正量 (HCE) 指标 (章节5) 来评估预测结果的质量。

讨论。表1和图2 (左) 说明了计算的度量就平均 IPQ 而言，本文的 DIS5K 比 SOD 数据集要复杂 20 倍 (最大达 50 倍)。尽管 CHAMELEON、COD10K、BIG、COIFT 和 ThinObject5K 相比 SOD 数据集有更高的平均 IPQ，但它们仍然比本文的数据集要简单得多。DIS5K 的平均 C_{num} 和标准差均在 100

和 400 以上。这表明 DIS5K 中的对象包含由多个轮廓组成的更精细的结构。DIS5K 的平均 P_{num} 在 1400 以上，比 HR-SOD 和合成的 ThinObject5K 分别高出近 5 倍和 3 倍。这三种复杂性度量是互补的，可以提供对对象复杂性的全面分析。表1中较大的标准差表明，从不同角度来看，DIS5K 的差异很大。更多结果请参考补充材料。图3-a 展示了来自 DUT-OMRON 数据集的观景台样本。本文的 DIS5K 中也包含了类似的对象 (b)，而具有更高的标注精度和结构复杂性。图3-c 为 COD10K 的样本，其相对于 SOD 数据集的结构复杂性较高的部分原因是遮挡也被标注了，而不是目标本身的结构复杂性。图3-d 显示了来自合成的 ThinObject5K 数据集的一组杠铃漂浮在空中的示例。图像合成是图像抠图中生成训练集的常用方法 [89,94]。但合成的图像通常与真实的图像不同，这导致了预测结果具有偏见。图3-e & f 展示了本文 DIS5K 类内结构复杂性的丰富多样。

3.3 数据集拆分

本文将 DIS5K 中的 5470 张图像分成三个子集: DIS-TR(3000)、DIS-VD(470) 和 DIS-TE(2000) 用于训练、验证和测试。DIS-TR 与 DIS-VD、DIS-TE 中包含的类别总体一致。由于本文的数据集中对象形状和结构复杂性的多样性，本文将 2000 张 DIS-TE 图像进一步按形状复杂性升序划分为四个子集，以便进行更全面的评估。具体来说，本文首先将 2000 幅测试图像按照其结构复杂性 IPQ 和边界复杂性 P_{num} 的乘积 ($IPQ \times P_{num}$) 升序排序。然后，将 DIS-TE 分为 4 个子集 (即, DIS-TE1∼DIS-TE4)，每个子集包含 500 张图像，代表 4 个测试难度等级。

4 本文的 IS-Net 基准

概述。如图4所示，本文的 IS-Net 由一个真值 (GT) 编码器、一个图像分割组件和一个本文新提出的中间监督策略组成。**GT 编码器** (27.7 MB) 用于将 GT 掩码编码到高维空间，然后用于对分割组件进行中间监督。同时，**图像分割组件** (176.6 MB) 被期望在可承受的内存和时间成本下，具有捕获精细结构并处理大尺寸（比如 1024×1024）输入的能力。在后续的实验中，本文选择 U2-Net [66] 作为图像分割组件，因为它对精细结构的捕捉能力较强。请注意，其他分割模型如 transformer 骨干网络也与本文的策略兼容。**技术细节**。 U^2-Net 最初是为小尺寸 (320×320) 的 SOD 图像设计的。由于其 GPU 内存成本，它不能直接用于处理大尺寸 (比如 1024×1024) 的输

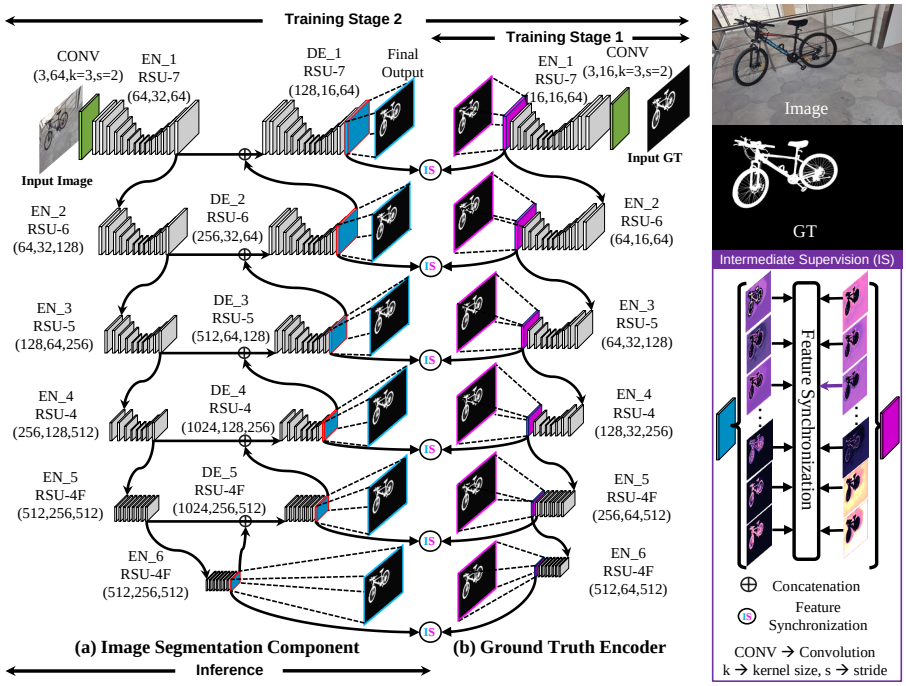


图 4: 本文的 IS-Net: (a) 表示图像分割组件、(b) 表示建立在中间监督 (IS) 组件上的真值编码器。

入。本文采用 U^2-Net 的结构，在其第一个编码器阶段之前增加一个输入卷积层。输入卷积层设为普通卷积层，核大小为 3×3，步长为 2。给定一个形状为 $I^{1024\times1024\times3}$ 的输入图像，输入卷积层首先将其转换为一个特征映射 $f^{512\times512\times64}$，然后该特征映射直接馈送到原始的 U^2-Net 中，相应的输入通道被更改为 64。与直接将 $I^{1024\times1024\times3}$ 输入到 U^2-Net 相比，输入卷积层在保持特征通道空间信息的同时，有助于整个网络减少总体 GPU 内存开销的四分之三。

4.1 中间监督

DIS 可以看作是分割模型中从图像域 $\mathcal{I} \in \mathbb{R}^{H\times W\times3}$ 到分割 GT 域 $\mathcal{G} \in \mathbb{R}^{H\times W\times1}$ 的映射: $\mathcal{G} = F(\theta, \mathcal{I})$, 其中 F 表示使用可学习权重 θ 将输入从图像映射到掩码域的模型。大多数模型在训练集上容易过度拟合。因此，深度监督被提出来对给定的深度网络 [41] 的中间输出进行监督。在文献 [66,88] 中，密集监督通常应用于侧输出，侧输出是通过卷积特定深度层的最后一层特征

图 5: 由本文的 GT 编码器的 EN_2 级的最后一层产生的特征图。“21”、“23”、“29” 和 “37” 为特征图中对应通道的索引 (以 1 开头)。

图而产生的单通道概率图。然而，将高维特征转换为单通道概率图本质上是一种降维操作，不可避免地会丢失关键线索。

为了避免这一问题，本文提出了一种新的中间监督训练策略。给定输入图像 $I^{H\times W\times 3}$ 及其对应的分割掩码 $G^{W\times H\times 1}$，本文首先使用一个轻量级的深度模型 F_{gt}，通过 “过拟合” 训练真值来训练一个自监督的 GT 编码器来提取高维特征，如图4-b 所示，$\underset{\theta_{gt}}{\operatorname{argmin}} \sum_{d=1}^{D} BCE(F_{gt}(\theta_{gt}, G)_d, G)$，其中 θ_{gt} 表示模型权重，BCE 表示二值交叉熵损失，D 表示中间特征图的数量。

获得 GT 编码器 F_{gt} 后，为了生成 “真值” 高维中间深度特征，冻结其权值 θ_{gt}。该操作主要通过: $f_D^G = F_{gt}^-(\theta_{gt}, G), D = \{1, 2, 3, 4, 5, 6\}$，其中 F_{gt}^- 用于生成概率图，表示不含最后一层卷积层的 F_{gt}。F_{gt}^- 被用来监督那些来自分割模型 F_{sg} 所对应的那些特征 f_D^I。在图像分割组件 F_{sg}(图4-a) 中，在生成概率图之前，图像 I 被转换为一组高维中间特征图 f_D^I。每个特征图 f_d^I 与其对应的 GT 中间特征图 f_d^G: $f_D^I = F_{sg}^-(\theta_{sg}, I), D = \{1, 2, 3, 4, 5, 6\}$ 具有相同的维度，其中 θ_{sg} 表示分割模型的权值。接着，中间监督通过下列高维特征一致性损失，对深度中间特征进行特征同步: $L_{\text{fs}} = \sum_{d=1}^{D} \lambda_d^{fs} \left\| f_d^I - f_d^G \right\|^2$，其中 λ_d^{fs} 为各 FS 损失的权重。分割模型 F_{sg} 的训练过程可以表述为如下优化问题: $\underset{\theta_{sg}}{\operatorname{argmin}}(L_{\text{fs}} + L_{\text{sg}})$，其中 L_{sg} 表示 F_{sg} 侧边输出的 BCE 损失: $L_{\text{sg}} = \sum_{d=1}^{D} \lambda_d^{sg} BCE(F_{sg}(\theta_{sg}, I), G)$，其中 λ_d^{sg} 表示加权各侧边输出损失的超参数。

图5为 GT 编码器在图4中阶段 2 的特征图 EN_2。我们可以看到输入掩码的多样化特征被编码到不同的通道中。例如 21^{st} 通道编码了接近原始掩码的细结构和大结构。23^{rd}、29^{th} 和 37^{th} 通道分别为中等尺寸结构 (车座和车轮)、精细结构 (刹车索和辐条) 和大尺寸区域 (自行车整体形状)。GT

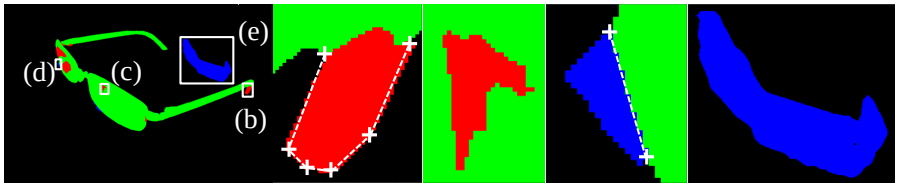
的这些多样性特征可提供更强的规整性和更全面的监督，以减少过度拟合的风险。

5 本文的 HCE 指标

给定一个预测分割概率 $P \in \mathbb{R}^{W\times H\times 1}$ 及其对应的 GT 掩码 $G \in \mathbb{R}^{W\times H\times 1}$，现有的指标，如 IoU、边界 IoU [10]、F-measure[2] [2]、边界 F-measure [2] 和 MAE [61]，一般通过基于 P 和 G 之间的数学或认知一致性 (或不一致性) 计算得分来评价预测结果 P 的质量。换句话说，这些指标描述了 P 和 G 之间的 “差距” 有多显著。然而，在许多应用中，评估填补 “缺口” 的成本比衡量 “缺口” 的大小更重要。

因此，本文提出了一种新的评估指标，人工校正量 (HCE)，它衡量了在现实应用中校正错误预测结果以满足特定的精度要求所需的人力。根据本文的标注经验，常用的操作主要有两种: (1) 沿目标边界点的选取，形成多边形和 (2) 基于区域内相似像素强度的区域选择。这两个操作都对应于一次鼠标点击。因此，这里的 HCE 是由鼠标点击次数来量化的。为了纠正错误的预测掩模，操作人员需要沿着错误预测目标的边界或区域手动采样主导点，以纠正假阳性 (FP) 和假阴性 (FN) 区域。如图6所示，FNs 和 FPs 根据其相邻区域可分为 $\mathrm{FN_N}$(N=TN+FP)、$\mathrm{FN_{TP}}$、$\mathrm{FP_P}$(P=TP+FN) 和 $\mathrm{FP_{TN}}$ 两类。为了校正 $\mathrm{FN_N}$ 区域，其与 TN 相邻的边界需要人工标记主导点 (图6-b)。同样，为了校正 $\mathrm{FP_P}$ 区域，本文只需要在 TP 区域附近标注其边界即可 (图6-d)。TP 包围的 $\mathrm{FN_{TP}}$ 区域 (图6-c) 和 TN 包围的 $\mathrm{FP_{TN}}$ 区域 (图6-e)，可以通过一键选择区域，轻松校正。因此，修正图6 (b-e) 中故障区域的 HCE 为 10(在 (b) 和 (d) 中需要点击 6 次和 2 次，在 (c) 中需要点击 1 次，在 (e) 中需要点击 1 次)。在评估阶段，根据 OpenCV findContours [77] 函数和连通区域标记算法 [26,87] 得到轮廓，分别用 DP 算法 [68] 逼近主导点选择操作和区域选择操作。

松弛 HCE。一些应用程序可以容忍某些较小的预测误差。因此，通过考虑误差公差 γ (HCE_γ) 可扩展 HCE。关键思想是通过使用腐蚀 [33] 和膨胀 [33] 来排除较小的 FP 和 FN 成分，从而松弛 FP 和 FN 区域。给定一个分割映射 P，它的 GT 掩码 G、误差容忍度 (比如 $\gamma = 5$，表示被忽略的小故障区域的大小)，DP 算法的 $epsilon$，HCE_γ 计算见算法 1。请注意，腐蚀操作可以去除 $P_{or}G$ 的所有薄和细组分。然而，一些薄的部件 (比如细电缆和网)



(a) 错误映射 (b) $\mathrm{FN_N}$ (c) $\mathrm{FN_{TP}}$ (d) $\mathrm{FP_P}$ (e) $\mathrm{FP_{TN}}$

图 6: 需要修复的故障区域。详见章节5。

Input: P, G, $\gamma = 5$, $epsilon = 2.0$
Output: HCE_γ

G_{ske} = skeletonize (G);
$P_{or}G$, TP = or (P, G), and (P,G);
FN, FP= (G - TP), (P - TP);
for $(i = 0; i \leq \gamma; i++)$ **do**
    $P_{or}G$ = erode ($P_{or}G$, $disk$ (1));
end
FN', FP' = and (FN,$P_{or}G$), and (FP,$P_{or}G$);
for $(i = 0; i \leq \gamma; i++)$ **do**
    FN' = dilate (FN', $disk$ (1));
    FN' = and (FN', not P);
    FP' = dilate (FP', $disk$ (1));
    FP' = and (FP', not G);
end
FN', FP' = and (FN, FN'), and (FP, FP');
FN' = or (FN', xor (G_{ske}, and (TP, G_{ske})));
HCE_γ = compute_HCE (FN', FP', TP, epsilon)

Algorithm 1: 松弛 HCE。

在描述目标时是至关重要的，因此需要保留它们。为此，[96] 提取了 GT 掩模的骨架，并与松弛的 FN' 掩模相结合，以保留这些结构。

6 DIS5K 基准

如前所述，本文的 DIS5K 从头开始构建，覆盖了具有非常不同几何结构和图像特征的高度多样化的对象。最重要的原因之一是排除现有数据集可能的偏差 (特定的图像或物体特征)。因此，它的多样性 (比如分辨率、图像特征、对象复杂性和标记精度) 和分布与现有数据集不同。为了提供一个公平的比较，所有模型分别在 DIS-TR、DIS-VD 和 DIS-TE 上进行训练、验证和测试。目前尚未进行跨数据集评测 [81]，主要原因是它们的标注精度与本文的不一致。

度量指标。为了提供相对全面和无偏的评估，本文采用 6 个不同的度量，包括最大 F-measure ($F_\beta^{mx} \uparrow$) [2]、加权 F-measure ($F_\beta^{w} \uparrow$) [52]、平均绝对误差

表 2: 对 DIS5K 验证和测试集进行定量评估。R = ResNet [35]。R2 = Res2Net [28]。S-813 = STDC813 [24], E-B1 = EffinetB1 [79]。

数据集	指标	UNet [71]	BASNet [67]	GateNet [101]	F^{3}Net [86]	GCPANet [8]	U^{2}Net [66]	SINetV2 [21]	PFNet [54]	PSPNet [99]	DLV3+ [6]	HRNet [82]	BSV1 [93]	ICNet [98]	MBV3 [36]	STDC [24]	HySM [58]	**IS-Net**
属性	骨干网络	-	R-34	R-50	R-50	R-50	-	R2-50	R-50	R-50	R-50	-	R-18	R-18	MBV3	S-813	E-B1	-
	模型大小 (MB)	121.4	348.6	515.0	102.6	268.7	176.3	108.5	186.6	196.1	161.8	264.4	47.6	46.5	21.5	48.4	49.6	176.6
	时间 (ms)	3.87	10.71	12.69	14.23	11.04	19.73	18.69	17.16	8.08	8.68	40.5	6.07	4.93	8.86	6.17	24.06	19.49
	输入大小	512^2	320^2	384^2	352^2	320^2	320^2	352^2	416^2	512^2	513^2	1024^2	1024x2048	1024x2048	1024^2	512x1024	512x1024	1024^2
DIS-VD	$maxF_\beta \uparrow$	0.692	0.731	0.678	0.685	0.648	0.748	0.665	0.691	0.691	0.660	0.726	0.662	0.697	0.714	0.696	0.734	**0.791**
	$F_\beta^w \uparrow$	0.586	0.641	0.574	0.595	0.542	0.656	0.584	0.604	0.603	0.568	0.641	0.548	0.609	0.642	0.613	0.640	**0.717**
	$M \downarrow$	0.113	0.094	0.110	0.107	0.118	0.090	0.110	0.106	0.102	0.114	0.095	0.116	0.102	0.092	0.103	0.096	**0.074**
	$S_\alpha \uparrow$	0.745	0.768	0.723	0.733	0.718	0.781	0.727	0.740	0.744	0.716	0.767	0.728	0.747	0.758	0.740	0.773	**0.813**
	$E_\phi^m \uparrow$	0.785	0.816	0.783	0.800	0.765	0.823	0.798	0.811	0.802	0.796	0.824	0.767	0.811	0.841	0.817	0.814	**0.856**
	$HCE_\gamma \downarrow$	1337	1402	1493	1567	1555	1413	1568	1606	1588	1520	1560	1660	1503	1625	1598	1324	**1116**
DIS-TE1	$maxF_\beta \uparrow$	0.625	0.688	0.620	0.640	0.598	0.694	0.644	0.646	0.645	0.601	0.668	0.595	0.631	0.669	0.648	0.695	**0.740**
	$F_\beta^w \uparrow$	0.514	0.595	0.517	0.549	0.495	0.601	0.558	0.552	0.557	0.506	0.579	0.474	0.535	0.595	0.562	0.597	**0.662**
	$M \downarrow$	0.106	0.084	0.099	0.095	0.103	0.083	0.094	0.094	0.089	0.102	0.088	0.108	0.095	0.083	0.090	0.082	**0.074**
	$S_\alpha \uparrow$	0.716	0.754	0.701	0.721	0.705	0.760	0.727	0.722	0.725	0.694	0.742	0.695	0.716	0.740	0.723	0.761	**0.787**
	$E_\phi^m \uparrow$	0.750	0.801	0.766	0.783	0.750	0.801	0.791	0.786	0.791	0.772	0.797	0.741	0.784	0.818	0.798	0.803	**0.820**
	$HCE_\gamma \downarrow$	233	220	230	244	271	224	274	253	267	234	262	288	234	274	249	205	**149**
DIS-TE2	$maxF_\beta \uparrow$	0.703	0.755	0.702	0.712	0.673	0.756	0.700	0.720	0.724	0.681	0.747	0.680	0.716	0.743	0.720	0.759	**0.799**
	$F_\beta^w \uparrow$	0.597	0.668	0.598	0.620	0.570	0.668	0.618	0.633	0.636	0.587	0.664	0.564	0.627	0.672	0.636	0.667	**0.728**
	$M \downarrow$	0.107	0.084	0.102	0.097	0.109	0.085	0.099	0.096	0.092	0.105	0.087	0.111	0.095	0.083	0.092	0.085	**0.070**
	$S_\alpha \uparrow$	0.755	0.786	0.737	0.755	0.735	0.788	0.753	0.761	0.763	0.729	0.784	0.740	0.759	0.777	0.759	0.794	**0.823**
	$E_\phi^m \uparrow$	0.796	0.836	0.804	0.820	0.786	0.833	0.823	0.829	0.828	0.813	0.840	0.781	0.826	0.856	0.834	0.832	**0.858**
	$HCE_\gamma \downarrow$	474	480	501	542	574	490	593	567	586	516	555	621	512	600	556	451	**340**
DIS-TE3	$maxF_\beta \uparrow$	0.748	0.785	0.726	0.743	0.699	0.798	0.730	0.751	0.747	0.717	0.784	0.710	0.752	0.772	0.745	0.792	**0.830**
	$F_\beta^w \uparrow$	0.644	0.696	0.620	0.656	0.590	0.707	0.641	0.664	0.657	0.623	0.700	0.595	0.664	0.702	0.662	0.701	**0.758**
	$M \downarrow$	0.098	0.083	0.103	0.092	0.109	0.079	0.096	0.092	0.092	0.102	0.080	0.109	0.091	0.078	0.090	0.079	**0.064**
	$S_\alpha \uparrow$	0.780	0.798	0.747	0.773	0.748	0.809	0.766	0.777	0.774	0.749	0.805	0.757	0.780	0.794	0.771	0.811	**0.836**
	$E_\phi^m \uparrow$	0.827	0.856	0.815	0.848	0.801	0.858	0.849	0.854	0.843	0.833	0.869	0.801	0.852	0.880	0.855	0.857	**0.883**
	$HCE_\gamma \downarrow$	883	948	972	1059	1058	965	1096	1082	1111	999	1049	1146	1001	1136	1081	887	**687**
DIS-TE4	$maxF_\beta \uparrow$	0.759	0.780	0.729	0.721	0.670	0.795	0.699	0.731	0.725	0.715	0.772	0.710	0.749	0.736	0.731	0.782	**0.827**
	$F_\beta^w \uparrow$	0.659	0.693	0.625	0.633	0.559	0.705	0.616	0.647	0.630	0.621	0.687	0.598	0.663	0.664	0.652	0.693	**0.753**
	$M \downarrow$	0.102	0.091	0.109	0.107	0.127	0.087	0.113	0.107	0.107	0.111	0.092	0.114	0.099	0.098	0.102	0.091	**0.072**
	$S_\alpha \uparrow$	0.784	0.794	0.743	0.752	0.723	0.807	0.744	0.763	0.758	0.744	0.792	0.755	0.776	0.770	0.762	0.802	**0.830**
	$E_\phi^m \uparrow$	0.821	0.848	0.803	0.825	0.767	0.847	0.824	0.838	0.815	0.820	0.854	0.788	0.837	0.848	0.841	0.842	**0.870**
	$HCE_\gamma \downarrow$	3218	3601	3654	3760	3678	3653	3683	3803	3806	3709	3864	3999	3690	3817	3819	3331	**2888**
Overall DIS-TE (1-4)	$maxF_\beta \uparrow$	0.708	0.752	0.694	0.704	0.660	0.761	0.693	0.712	0.710	0.678	0.743	0.674	0.711	0.729	0.710	0.757	**0.799**
	$F_\beta^w \uparrow$	0.603	0.663	0.590	0.614	0.554	0.670	0.608	0.624	0.620	0.584	0.658	0.558	0.622	0.658	0.628	0.665	**0.726**
	$M \downarrow$	0.103	0.086	0.103	0.098	0.112	0.083	0.101	0.097	0.095	0.105	0.087	0.110	0.095	0.085	0.094	0.084	**0.070**
	$S_\alpha \uparrow$	0.759	0.783	0.732	0.750	0.728	0.791	0.747	0.756	0.755	0.729	0.781	0.737	0.758	0.770	0.754	0.792	**0.819**
	$E_\phi^m \uparrow$	0.798	0.835	0.797	0.819	0.776	0.835	0.822	0.827	0.819	0.810	0.840	0.778	0.825	0.850	0.832	0.834	**0.858**
	$HCE_\gamma \downarrow$	1202	1313	1339	1401	1395	1333	1411	1427	1442	1365	1432	1513	1359	1457	1426	1218	**1016**

($M \downarrow$) [61]、结构度量 ($S_\alpha \uparrow$) [19]、平均增强对准度量 ($E_\phi^m \uparrow$) [20, 22] 和本文的人工矫正量 ($HCE\gamma \downarrow$) 从不同的角度来评估模型性能。

对比算法。本文将 IS-Net 与 16 个设计用于不同分割任务的主流模型进行了比较，包括 (i) 医学图像分割模型 U-Net [71];(ii) 显著目标检测模型，如 BASNet [67]、GateNet [101]、F^{3}Net [86]、GCPA [8] 和 U2-Net [66];(iii) 针对 COD 设计的 SINet-V2 [21] 和 PFNet [54] 等; 语义分割模型: PSPNet [99]、DeepLab-V3+ [6] 和 HRNet [82];(v) 实时语义分割模型: BiSeNetV1 [93]、ICNet [98]、MobileNet-V3-Large [36]、STDC [25] 和 HyperSegM [58]。所有模型使用 DIS-TR 集进行重新训练 (Tesla V100 或 RTX A6000)，表2中的预测时间成本均在 RTX A6000 上进行测试。

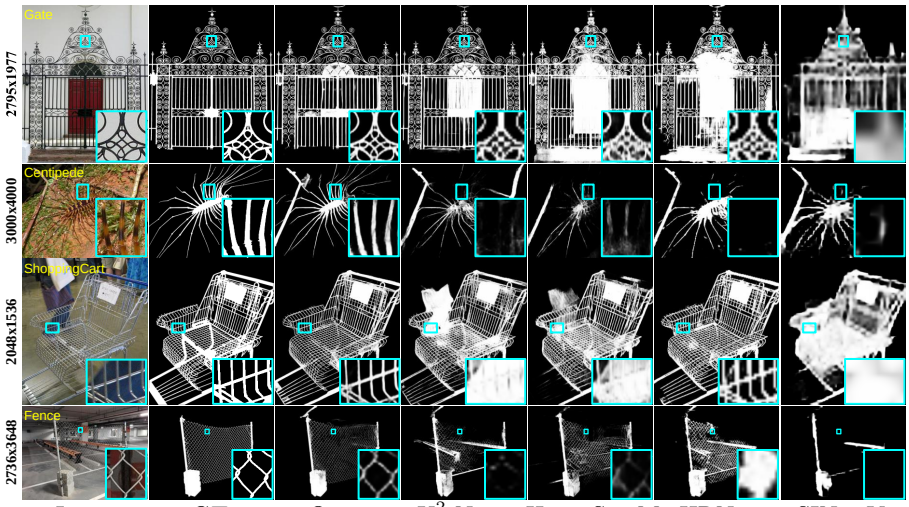


图 7: IS-Net 与四个基线的定性比较。

6.1 定量评价

与 16 种 SOTA 模型相比，本文的 IS-Net 在所有指标上都达到了最具竞争力的性能 (见表2)。本文观察到，不同模型的性能可能与模型输入的空间大小及其特征图有关。大多数分割模型都引入了分类主干网络来构建它们的编码器-解码器体系结构。然而，像 ResNet-50 [35] 这样的主干网络首先是输入卷积层 (步长为 2)，然后是池化操作 (步长为 2)，以减少特征映射的空间大小，这导致了大量空间信息的丢失和性能的显著下降。当待分割目标的形状接近凸时，退化不明显。然而，DIS5K 中很多物体都是非凸的，它们的结构非常复杂精细。它要求模型尽可能地保留空间信息，这对大多数模型都是一个挑战。

6.2 定性评价

图7展示了本文的方法与四个 SOTA 基准之间的定性比较。本文的模型在不同的场景中，无论是显著 (大门)、伪装 (蜈蚣)、结构复杂 (购物车) 还是细小的 (栅栏) 物体，都取得了很好的效果，体现了本文的 IS-Net 基准的泛化能力。

6.3 消融实验

为了验证本文新提出的中间监督策略适配在最近 SOTA 模型如 U^2-Net 上的有效性，本文进行了全面的消融研究。

表 3: 在 DIS-VD 数据子集上进行消融研究。

Settings	$F_\beta^{mx}\uparrow$	$F_\beta^{w}\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi^m\uparrow$	$HCE_\gamma\downarrow$
U^2-Net 320^2 (baseline)	.748	.656	.090	.781	.823	1413
U^2-Net 512^2	.769	.677	.085	.789	.826	1146
U^2-Net 1024^2	.764	.667	.088	.792	.820	1085
U^2-Net 1024^2 (**Adp**)	**.776**	**.695**	**.080**	**.804**	**.844**	**1076**
Adp+Last-1(L_2)	.777	.695	.080	.799	.840	1115
Adp+Last-2(L_2)	.778	.704	.079	.803	.847	**1049**
Adp+Last-3(L_2)	.788	.708	.079	**.812**	.845	1078
Adp+Last-4(L_2)	.782	.703	.079	.807	.849	1063
Adp+Last-5(L_2)	.788	**.715**	**.074**	.811	**.853**	1059
Adp+Last-6(L_2)	**.790**	.710	**.074**	.810	.852	1056
Adp+Last-6(KL)	.770	.684	.084	.794	.837	1092
Adp+Last-6(L_1)	.770	.686	.080	.797	.837	1144
Adp+Last-6(L_2) (shared outconv)	.745	.646	.094	.779	.813	1191
Adp+Last-6(L_2,sd(1))	.786	.706	.076	.807	.844	1086
Adp+Last-6(L_2,sd(58))	.790	.709	.078	.812	.848	**1085**
Adp+Last-6(L_2,sd(472))	.790	.712	.075	.812	.852	1071
Adp+Last-6(L_2,sd(5289)) (**IS-Net**)	**.791**	**.717**	**.074**	**.813**	**.856**	1116

输入大小。从表3可以看出，较大的输入大小可以提高 U^2-Net 的精度。但是，这会增加 GPU 内存成本。所以当输入大小为 1024×1024 时，需要减少批量大小 (Tesla V100 为 3, 32 GB)，这会降低性能。本文的简单而有效的变体 (即, Adp, 第 4 行) 解决了这个内存问题并提高了性能。

不同解码器阶段的监督设定。在表3中 last-S 表示对最后 S 解码器进行中间监督。如图所示，在 Last-6 阶段应用中间监督可以获得相对更好的性能，这是本文的默认设置。

不同的损失函数。不同损失函数的结果表明，L_2 比 KL 散度和 $L1$ 好。此外，共享 GT 编码器和分割解码器的 “outconvs” (将深度特征映射转换为分割概率映射) 也会带来负面影响。

随机种子。为了研究随机初始化权值的影响，本文用不同的随机种子初始化权值对同一个 GT 编码器进行多次训练。可以看出，虽然不同的随机种子产生的精度不同，但它们的变化很小，都优于没有中间监督策略训练的模型 (U^2-Net 和 Adp)。由于种子 5289 的模型在六个总体指标中有五个排名第一，本文使用这个模型作为本文的基线模型 IS-Net。

7 结论

本文从应用和研究两方面对高精度图像二类分割 (DIS) 任务进行了系统的研究。为了证明该任务可解，本文建立了一个新的具有挑战性的 **DIS5K** 数据集，引入了一个简单有效的中间监督网络，称为 IS-Net，以实现高质量的实时分割结果，并设计了一个新的人工矫正量 (**HCE**) 指标，考虑物体形状的复杂性。通过充分的消融研究和全面的基准测试，本文展示了新提出的 DIS 分割挑战任务是可解的。

参考文献

1. Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index, accessed: 2021-09-21
2. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR (2009)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI **39**(12), 2481–2495 (2017)
4. Birsan, T., Tiba, D.: One hundred years since the introduction of the set distance by dimitrie pompeiu. In: IFIP SMO (2005)
5. Blumberg, H.: Hausdorff's Grundzüge der Mengenlehre. Bulletin of the American Mathematical Society, 27 (3): 116–129, American (1920)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
7. Chen, S., Ma, X., Lu, Y., Hsu, D.: Ab initio particle-based object manipulation. In: Shell, D.A., Toussaint, M., Hsieh, M.A. (eds.) RSS (2021)
8. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: AAAI (2020)
9. Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A.: Boundary IoU: Improving object-centric image segmentation evaluation. In: CVPR (2021)
10. Cheng, B., Girshick, R.B., Dollár, P., Berg, A.C., Kirillov, A.: Boundary iou: Improving object-centric image segmentation evaluation. In: CVPR (2021)
11. Cheng, H.K., Chung, J., Tai, Y.W., Tang, C.K.: Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In: CVPR (2020)
12. Cheng, M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.: Global contrast based salient region detection. IEEE TPAMI **37**(3), 569–582 (2015)
13. Chinchor, N.: MUC-4 evaluation metrics. In: MUC (1992)
14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)

16. Ehrig, M., Euzenat, J.: Relaxed precision and recall for ontology matching. In: K-CapW (2005)
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010)
18. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: ECCV (2018)
19. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV (2017)
20. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: IJCAI (2018)
21. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE TPAMI (2021)
22. Fan, D.P., Ji, G.P., Qin, X., Cheng, M.M.: Cognitive vision inspired object segmentation metric and loss function. SSI **6** (2021)
23. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: CVPR (2020)
24. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: CVPR (2021)
25. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: CVPR (2021)
26. Fiorio, C., Gustedt, J.: Two linear time union-find strategies for image processing. TCS **154**(2), 165–181 (1996)
27. Freixenet, J., Muñoz, X., Raba, D., Martí, J., Cufí, X.: Yet another survey on image segmentation: Region and boundary information integration. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV (2002)
28. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE TPAMI **43**(2), 652–662 (2019)
29. Girshick, R.: Fast r-cnn. In: ICCV (2015)
30. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
31. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE TPAMI **34**(10), 1915–1926 (2012)
32. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`

33. Haralick, R.M., Sternberg, S.R., Zhuang, X.: Image analysis using mathematical morphology. IEEE TPAMI **PAMI-9**(4), 532–550 (1987)

34. Hausdorff, F.: Grundzüge der Mengenlehre. Leipzig: Veit, ISBN 978-0-8284-0061-9Reprinted by Chelsea Publishing Company in 1949, Germany (1914)

35. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

36. Howard, A., Pang, R., Adam, H., Le, Q.V., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y.: Searching for mobilenetv3. In: ECCV (2019)

37. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: CVPR (2020)

38. Ke, Z., Li, K., Zhou, Y., Wu, Q., Mao, X., Yan, Q., Lau, R.W.: Is a green screen really necessary for real-time portrait matting? ArXiv **abs/2011.11961** (2020)

39. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012)

40. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. CVIU **184**, 45–56 (2019)

41. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS (2015)

42. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR (2015)

43. Li, H., Xiong, P., Fan, H., Sun, J.: Dfanet: Deep feature aggregation for real-time semantic segmentation. In: CVPR (2019)

44. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR (2014)

45. Liew, J.H., Cohen, S., Price, B., Mai, L., Feng, J.: Deep interactive thin object selection. In: WACV (2021)

46. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. CoRR **abs/2108.11515** (2021)

47. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)

48. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. IEEE TPAMI **33**(2), 353–367 (2011)

49. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)

50. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)
51. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: CVPR (2021)
52. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps. CVPR (2014)
53. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE TPAMI **26**(5), 530–549 (2004)
54. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: CVPR (2021)
55. Mnih, V.: Machine Learning for Aerial Image Labeling. Ph.D. thesis, University of Toronto (2013)
56. Mnih, V., Hinton, G.E.: Learning to detect roads in high-resolution aerial images. In: ECCV (2010)
57. Movahedi, V., Elder, J.H.: Design and perceptual validation of performance measures for salient object segmentation. In: CVPRW (2010)
58. Nirkin, Y., Wolf, L., Hassner, T.: Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. arXiv preprint arXiv:2012.11582 (2020)
59. Orsic, M., Kreso, I., Bevandic, P., Segvic, S.: In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: CVPR (2019)
60. Osserman, R.: The isoperimetric inequality. BAM **84**(6), 1182–1238 (1978)
61. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR (2012)
62. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
63. Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Lin, Z., Torr, P., Jia, J.: Open-world entity segmentation. arXiv preprint arXiv:2107.14228 (2021)
64. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Gool, L.V.: Highly accurate dichotomous image segmentation. In: ECCV (2022)
65. Qin, X., Fan, D.P., Huang, C., Diagne, C., Zhang, Z., Sant'Anna, A.C., Suàrez, A., Jagersand, M., Shao, L.: Boundary-aware segmentation network for mobile and web applications. arXiv preprint arXiv:2101.04704 (2021)

66. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. PR **106**, 107404 (2020)

67. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: CVPR (2019)

68. Ramer, U.: An iterative procedure for the polygonal approximation of plane curves. CGIP **1**(3), 244–256 (1972)

69. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS (2015)

70. van Rijsbergen, C.J.: Information retrieval. London:Butterworths, `1979.http://www.dcs.gla.ac.uk/Keith/Preface.html` (1979)

71. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)

72. Saito, S., Yamashita, T., Aoki, Y.: Multiple object extraction from aerial imagery with convolutional neural networks. EI **2016**(10), 1–9 (2016)

73. Shen, X., Hertzmann, A., Jia, J., Paris, S., Price, B., Shechtman, E., Sachs, I.: Automatic portrait segmentation for image stylization. In: CGF (2016)

74. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)

75. Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., Kozieł, P.: Animal camouflage analysis: Chameleon database. Unpublished Manuscript (2018)

76. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR **15**(1), 1929–1958 (2014)

77. Suzuki, S., Abe, K.: Topological structural analysis of digitized binary images by border following. CVGIP **30**(1), 32–46 (1985)

78. Sørensen, T.J.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. København, I kommission hos E. Munksgaard, Denmark (1948)

79. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. pp. 6105–6114 (2019)

80. Tang, L., Li, B., Zhong, Y., Ding, S., Song, M.: Disentangled high quality salient object detection. In: ICCV (2021)

81. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
82. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI (2019)
83. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017)
84. Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., Borji, A.: Detect globally, refine locally: A novel approach to saliency detection. In: CVPR (2018)
85. Watson, A.B.: Perimetric complexity of binary digital images. Math J **14**, 1–40 (2012)
86. Wei, J., Wang, S., Huang, Q.: F^3net: Fusion, feedback and focus for salient object detection. In: AAAI (2020)
87. Wu, K., Otoo, E.J., Shoshani, A.: Optimizing connected component labeling algorithms. In: Fitzpatrick, J.M., Reinhardt, J.M. (eds.) MI (2005)
88. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
89. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: CVPR (2017)
90. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR (2013)
91. Yang, C., Wang, Y., Zhang, J., Zhang, H., Lin, Z., Yuille, A.: Meticulous object segmentation. arXiv preprint arXiv:2012.07181 (2020)
92. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR (2013)
93. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV (2018)
94. Yu, H., Xu, N., Huang, Z., Zhou, Y., Shi, H.: High-resolution deep image matting. arXiv preprint arXiv:2009.06613 (2020)
95. Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. In: CVPR. pp. 7234–7243 (2019)
96. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Commun. ACM **27**(3), 236–239 (1984)
97. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. GRSL **15**(5), 749–753 (2018)
98. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: ECCV (2018)
99. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)

100. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: ICCV (2019)

101. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: A simple gated network for salient object detection. In: ECCV (2020)

102. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR (2021)

103. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE TPAMI **40**(6), 1452–1464 (2017)

104. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)