使用 TSP6K 数据集解析交通场景

Peng-Tao Jiang[1,2*] Yuqi Yang[1*] Yang Cao[3] Qibin Hou[1,4†] Ming-Ming Cheng[1,4] Chunhua Shen[2]

[1]VCIP, CS, Nankai University [2]Zhejiang University [3]HKUST [4]NKIARI, Shenzhen

pt.jiang@mail.nankai.edu.cn, yangyq2000@mail.nankai.edu.cn, andrewhoux@gmail.com

摘要

计算机视觉中的交通场景感知是实现智能城市的一项至关重要的任务。迄今为止，大多数现有数据集都集中于自动驾驶场景。我们观察到，在这些驾驶数据集上训练的模型在交通监控场景中往往产生让人不满意的结果。然而，由于缺乏特定的数据集，在改善交通监控场景理解这一方面投入的努力很少。为了填补这一空白，我们引入了一个专门的交通监控数据集，它被称为 *TSP6K*，其中包含来自交通监控场景的图像，并带有高质量的像素级以及实例级注释。*TSP6K* 数据集中交通参与者的数量是现有驾驶场景的几倍，捕捉到了更加拥挤的交通场景。我们对数据集进行了详细的分析，并全面评估了以前流行的场景解析、实例分割和无监督领域自适应的方法。此外，考虑到实例大小的巨大差异，我们提出了一种用于场景解析的细节优化解码器，该解码器可以根据所提出的 *TSP6K* 数据集恢复交通场景中不同语义区域的细节。我们用实验证明了其在解析交通监控场景中的有效性。代码和数据集可在以下网址获取：*https://github.com/PengtaoJiang/TSP6K*.

1. 引言

场景解析任务是一项经典且重要的计算机视觉任务，其目的是从给定的图像中分割出语义对象和内容。如今，随着大规模场景理解数据集的出现，如 ADE20K [101]、COCO-Stuff [5] 等，极大地促进了场景理解算法的发展 [21, 43, 52, 57, 83, 98, 100]. 许多如机器人导航 [18, 42] 和医疗诊断 [62] 的应用场景，都受益于先进的场景理解算法。作为场景理解的一个重要案例，交通场景理解专注于理解城市街道场景，其中最常出现的实例是人、车辆和交通标志。目前，已有许多大规模的公开交通场景数据集，例如 KITTI [25]、Cityscapes [17] 和 BDD100K [87]。得益于这些精细注释的数据集，最近的场景理解方法 [13, 30, 49, 51, 66, 78, 94, 102] 的分割性能也得到了显着提升。

这些交通数据集的一个特点是，它们大多是从驾驶平台（例如驾驶汽车）收集的，因此更适合自动驾驶场景。然而，交通监控场景却很少受到关注。交通监控场景通常由街道高挂（4.5-6 米）的拍摄平台进行拍摄，可以提供丰富的交通流信息 [44, 58]。高悬拍摄台通常能观察到比驾车者更多的交通参与者，尤其是在十字路口。我们观察到，或许是因为域差异，在这些现有交通数据集上训练的深度学习模型在解析交通监控场景时会取得较差的效果。尽管许多应用都需要分析交通监控场景，例如交通流分析 [44, 58]，但据我们所知，目前没有可用于促进此类研究的交通场景数据集。

为了推动对交通监控场景进行解析的研究，我们构建了一个专门用于交通场景分析的数据集，并在本文中给出。具体来说，我们会从不同地点的城市道路拍摄平台精心收集了大量的交通图像。为了保持数据集的多样性，我们收集了一天中不同时间的数百个交通场景的图像。为了对该数据集进行语义分割和实例分割，我们要求注释者使用高质量的语义和实例级标签对其进行精细注释。由于标注的人工成本昂贵，我们

*The first two authors contributed equally to this work. Part of this work was done when P.-T. Jiang was a postdoc researcher at Zhejiang University.

†Q. Hou is the corresponding author.

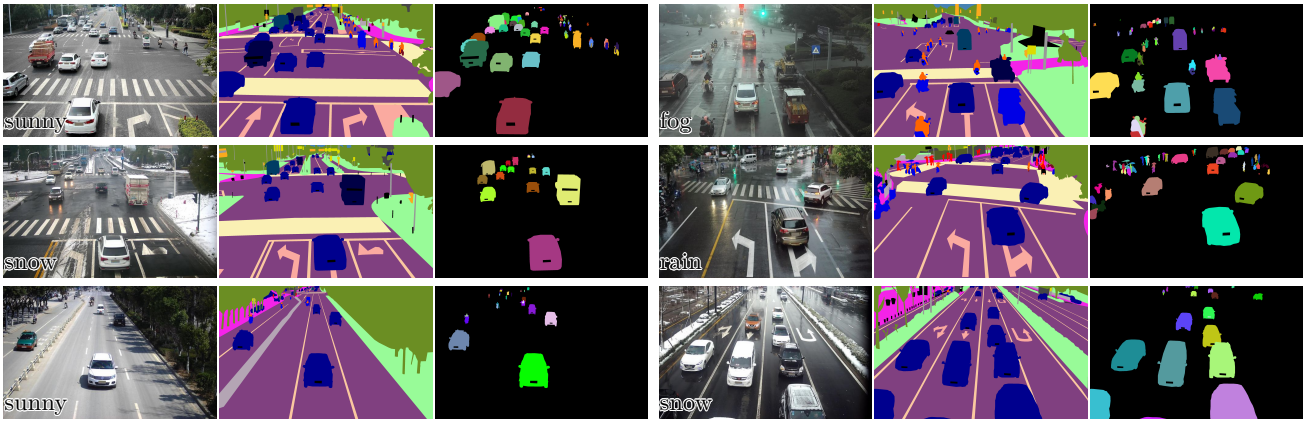


图 1. 从 TSP6K 数据集中随机挑选的样例。每张图像都对应其相应的语义标签和实例标签。我们已对车辆牌照进行了遮挡以保护隐私。

最终获得了 6,000 幅精细标注的交通图像。在 图 1 中，我们展示了一些交通图像及其相应的语义级和实例级标签。利用这些精细注释的标签，我们对交通监控场景进行了全面的研究。该数据集的特点概括如下：**1)** 最大的交通监控数据集，**2)** 更加拥挤的场景，**3)** 实例大小变化很大，**4)** 驾驶场景和监控场景之间的域差异很大。基于提出的 TSP6K 数据集，我们评估了一些经典的场景解析方法、实例分割方法和无监督域自迁移方法。我们分析并展示了不同方法在所提出的 TSP6K 数据集上的性能。

此外，我们提出了一种用于 TSP6K 上的图像分割的细节优化解码器。细节优化解码器使用编码器-解码器结构，并使用区域细化模块细化高分辨率特征。区域细化模块利用注意力机制并计算像素与每个区域标记之间的注意力。注意力机制进一步用于细化不同语义区域中的像素关系。我们以 SegNeXt [27] 的主干作为编码器，本文提出的细节优化解码器方法在 TSP6K 验证集上实现了 75.8% 的 mIoU 分数和 58.4% 的 iIoU 分数，分别比最先进的 SegNeXt 工作高出 1.2% 和 1.1%。总而言之，我们的主要贡献如下：

- 我们提出了一个专门用于研究交通监控场景解析任务的交通数据集，称为 TSP6K，它从城市道路拍摄平台收集涵盖各种场景的图像。我们还提供了语义标签和实例标签的像素级注释。
- 基于 TSP6K 数据集，我们评估了以前的场景解析方法和实例分割方法在交通监控场景上的性能。此外，TSP6K 数据集还可以作为评估交通监控应用的无监督领域自迁移方法的额外补充。
- 为了改进交通监控场景解析，我们提出了一个细节优化解码器，它学习区域标记以细化高分辨率特征上的不同区域。实验验证了所提出的解码器的有效性。

2. 相关工作

2.1. 场景解析数据集

具有完整像素注释的场景解析数据集常被用于训练和评估场景解析算法。作为早期数据集，PASCAL VOC 数据集 [22] 是在一项挑战中提出的，旨在解析每幅图像中 20 个精心挑选的类别的对象。后来，社区提出了更复杂、类别更多的数据集，例如 COCO [54] 和 ADE20K [101]。上述数据集中的场景涵盖范围很广。与这些数据集不同的是，还有一些数据集专注于特定场景，例如交通场景。目前存在许多交通场景解析数据集 [19, 39, 60, 64, 72, 90, 91]，例如 KITTI [25]，Cityscapes [17]，ACDC [65]，和 BDD100K [87]。这些交通解析数据集注释了交通场景中最常见的类别，例如交通标志、骑手和车辆等。基于这些精细注释的交通数据集，使用神经网络的方法在解析交通场景方面取得了巨大成功。

尽管上述数据集取得了成功，但我们发现这些数据集中的交通场景全部来自驾驶平台。在这些数据集上训练的模型在交通监控场景中往往表现不佳，而交通监控在交通流分析中起着重要作用。此外，监控场景通常比驾驶场景捕捉到更多的交通参与者。而我们提出的 TSP6K 数据集不同于驾驶数据集，旨在提高场景

解析模型在监控场景上的性能，可以看作是对现有交通数据集的补充。此外，Kirillov 等人 [46] 提出了一个包含 10 亿个掩码的大型分割数据集 SA-1B。SA-1B 也包含了一些监控交通图像。然而，SA-1B 中的分割掩码都是与类别无关的。与此相反，我们数据集中的分割掩码都是类别已知的。

2.2. 场景解析方法

卷积神经网络极大地促进了场景解析方法的发展。有代表性的是 Long 等人 [57] 首次提出的一种完全卷积网络 (FCN)，它可以为场景解析生成密集的预测。后来，一些方法，例如流行的 DeepLab [9, 10] 和 PSPNet [98]，得益于较大的感受野和多尺度特征，大大提高了性能。此外，还有一些方法 [3, 12, 13, 52]，利用编码器-解码器结构，利用高分辨率特征的细节来细化低分辨率粗预测。除了简单的卷积之外，研究人员 [36, 41, 89, 99] 发现，注意力机制 [29] 可以显著改善场景解析网络，因为它们能够对长距离依赖关系进行建模。此外，还有一些工作 [79, 85, 86, 95, 97] 探索实时场景解析算法，这些算法以有效的方式利用了自注意力。

最近，随着 Transformer 网络成功应用到图像识别 [20]，研究人员尝试将 Transformer 网络应用于分割任务 [14, 15, 66, 83, 100]。有趣的是，最近的一些研究 [27, 28, 35] 表明，卷积神经网络在场景解析任务上的表现优于基于 Transformer 的模型。在我们的数据集中，我们也观察到了类似的结果。SegNeXt [27] 在我们的 TSP6K 数据集上取得了最佳性能，并且使用的参数比其他工作更少。本研究提出的方法也采用了 SegNeXt 工作的主干网络。但与它不同的是，我们设计了一个细节优化解码器，它比 SegNeXt 中使用的解码器更适合处理高分辨率图像。

2.3. 实例分割方法

实例分割的目的是对同一类中的每个实例进行分割和区分。以前的实例分割方法大致可以分为两类，依赖边框的方法 [4, 7, 23, 31, 40, 55]，和无边框方法 [68, 75, 76, 82]。基于边框的方法首先检测目标对象的边界框，然后在框区域内进行二值分割。相比之下，无边框方法直接为每个实例生成实例掩码并并行对其进行分类。在本文中，我们为每个类别选择了几种方法，并在 TSP6K 上对它们进行了评估。

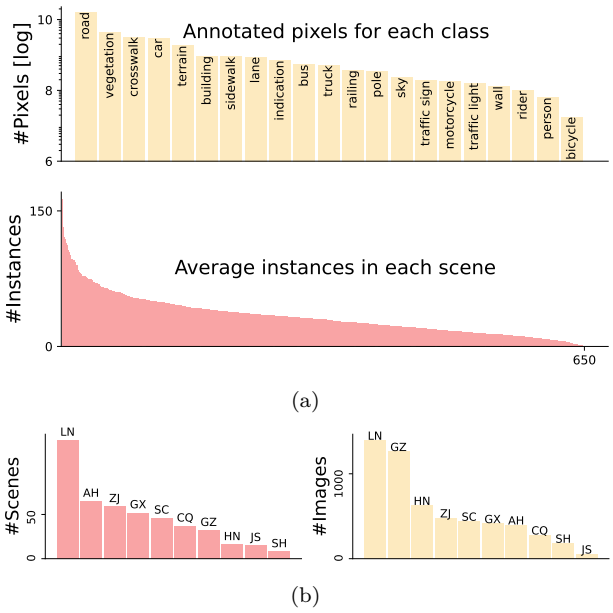


图 2. (a) TSP6K 数据集的类别和场景的信息。(b) 图像拍摄场景的地理位置分布。

2.4. 无监督域自迁移

无监督域自迁移 (UDA) 旨在对模型进行自适应，从一个域 (有分割标签) 迁移到一个新域 (没有) 细分标签)。近年来，许多用于场景解析的 UDA 方法 [33, 70, 73, 84] 为了解决域之间的差距被提出。UDA 方法主要分为两类：基于对抗训练的方法 [33, 34, 70, 71, 73, 74]，以及基于自训练的方法 [47, 50, 59, 77, 93, 103, 104]。基于对抗训练的方法试图对齐源域和目标域的特征表示或网络预测。而基于自训练的方法则为目标域生成伪掩码来训练分割网络。以往的无监督域迁移（UDA）方法通常通过从合成交通数据集（如 GTA5 [61] 或 SYNTHIA [63]）迁移到真实交通数据集（如 Cityscapes [17]）来进行评估。在本文中，我们评估了无监督域迁移（UDA）方法，通过从合成和真实驾驶场景（SYNTHIA [63]，Cityscapes [17]）迁移到监控场景（TSP6K）。

3. 数据集与分析

在本节中，我们介绍构建监测数据集的细节，并对提出的 TSP6K 数据集进行全面分析。

表 1. 不同交通场景解析数据集之间的比较。**Avg TP** 表示交通参与者的平均数量。**TP > 50** 表示包含超过 50 个交通参与者的图像数量。由于其他数据集中测试集的实例标签不可用，我们统计了训练集和验证集中的交通参与者数量。可以看出，与其他数据集相比，我们的 TSP6K 数据集包含更多超过 50 个交通参与者的交通图像。

Type	Datasets	Class	Weather	#Images	Resolution	Annotation	Avg TP	TP > 50	TP > 75	TP > 100
Driving	KITTI [25]	19	Good	7,481	1,241×375	Pixel&Inst	4.9	0	0	0
	Cityscapes [17]	19	Good	5,000	2,048×1,024	Pixel&Inst	18.8	54	10	4
	WildDash 2 [90]	26	Diverse	5,068	1920×1080	Pixel&Inst	9.0	12	4	2
	Mapillary [60]	65	Diverse	25,000	3,436×2,486	Pixel&Inst	12.3	102	15	3
	ACDC [65]	19	Diverse	3,142	1080×1920	Pixel&Inst	6.3	0	0	0
	BDD100K [87]	40	Good	10,000	1,280×720	Pixel&Inst	12.8	5	0	0
Monitoring	UrbanTracker [45]	7	Good	5(videos)	1,035×632	Box	3.7	0	0	0
	CityFlow [67]	1	Good	40(videos)	540×960	Box	2.0	-	-	-
	AAU RainSnow [1]	3	Diverse	22(videos)	640×480	Pixel	6.6	0	0	0
	TSP6K (ours)	21	Diverse	6,000	2,942×1,989	Pixel&Inst	42.0	1,227	367	73

表 2. 交通图像中交通参与者的统计数据。'#H.'（'#Humans'）和 '#V.'（'#Vehicles'）分别表示人数和车辆数。

Datasets	#Humans $[10^3]$	#Vehicles $[10^3]$	#H./images	#V./images
KITTI [25]	6.1	30.3	0.8	4.1
Cityscapes [17]	24.4	41.0	7.0	11.8
Mapillary [60]	6.7	17.8	3.4	8.9
Wilddash2 [90]	11.6	26.8	2.7	6.3
ACDC [65]	3.8	15.9	1.2	5.1
BDD100K [87]	11.7	90.3	1.5	11.3
TSP6K (ours)	64.0	188.2	10.7	31.3

3.1. 数据收集

研究交通监控场景的一个重要方面是数据。一旦我们为监控场景构建了数据集，社区研究人员就可以根据新的数据特征改进场景解析结果。为了方便研究，我们计划通过从不同街道上的高悬拍摄平台收集大量图像，建立一个专门用于研究交通监控场景的数据集。为了确保数据的多样性，采集地点和天气条件被高度重视。具体来说，我们收集了中国约 10 个省份的交通图像，超过 600 个场景。在 图 2(b) 中，我们展示了场景和图像的地理分布。由于人行横道和行人过街是交通场景的重要组成部分，经常发生拥堵和事故，因此我们保留了大部分包含人行横道的交通场景。此外，考虑到天气的多样性，我们选取了各种天气条件下的交通图像，包括晴天、阴天、雨天、雾天、雪天。最终我们选定了 6000 张交通图像。

3.2. 数据标注

数据采集完成后，我们开始对交通图像进行标注。完整的标注类别如图 2(a) 所示。具体而言，我们标注了 21 个类别，其中大部分类别与 Cityscapes [17] 中的类别定义相同。我们移除了数据集中未出现的类别 "火车"，并添加了三个新类别。由于道路指示对理解监控场景至关重要，我们要求标注人员为交通标注三个指示类别，分别是人行横道、行驶指示和车道线。此外，我们还为每个交通参与者标注了实例掩码。

类似于 Cityscapes [17] 的标注策略，交通图像也采用了从背景到前景的标注方式。为了保证标签的质量，我们设计了双重检查机制。具体而言，图像被分成 30 组，每组包含 200 张图像。当标注人员完成图像标注后，我们从 200 张图像中随机抽取 30% 的图像，以检查是否存在类别标注错误。如果在选定的图像中发现类别标注错误，我们要求标注人员检查该组中的所有图像，直到没有类别标注错误为止。

3.3. 数据划分

数据集按照 5:2:3 的比例划分为训练集、验证集和测试集。来自不同场景的图像被随机分配到不同的集合中。总共有 2999 张、1207 张和 1794 张图像分别用于训练集、验证集和测试集。

3.4. 数据分析

我们对比了 TSP6K 数据集与之前的交通数据集，在场景类型、实例密度、实例尺度变化和领域差异等方面进行比较。在表 1 和表 2 中，我们列出了不同交通数据集之间的比较。TSP6K 数据集的特点可以总结如下：

规模最大的交通监控数据集： 据我们所知，大多数之前的流行交通数据集主要集中在驾驶场景上。这些数据集的图像是从驾驶平台上收集的。还有几个数据集 [45, 67] 包括交通监控场景，如表 1所示。然而，这些数据集主要关注交通参与者的跟踪，仅提供了类别无关的边界框标注。与之不同的是，我们专注于交通监控场景的解析，并提供了语义和实例标注。与现有的交通监控数据集相比，TSP6K 包含了更多标注的语义类别，提供了实例分割，更高的图像分辨率，以及更多的图像数量。

更加拥挤的场景： 其中一个最重要的特点是，TSP6K 数据集包含了比驾驶数据集更为拥挤的图像。由于大多数交通场景是在交叉路口拍摄的，因此道路上的实例密度远高于驾驶场景。在 表 1中可以看到，驾驶数据集中很少有图像包含超过 50 个实例。相比之下，我们的 TSP6K 数据集有大量图像包含超过 50 个实例，这些图像占训练集和验证集中约 30%。此外，如表 2所示，TSP6K 中平均有 10.7 个人和 31.3 辆车，这一数量是其他驾驶数据集的几倍。此外，可以看出，现有的监控数据集的标注实例数量少于驾驶数据集。这主要是因为标注不完整，只对移动的车辆进行了标注。

实例尺度差异大： 在监控场景中，前景和背景中的实例大小差异非常大。如图 3(b) 所示，TSP6K 的实例尺寸范围比 Cityscapes 更广。此外，TSP6K 还包含比 Cityscapes 更多的小型交通实例。高悬平台通常具有比驾驶平台更广阔的视野。因此，它可以捕捉到更远处的更多内容。实例尺寸的巨大变化展示了真实的交通场景。

域差异大： TSP6K 与 Cityscapes/BDD100K 之间存在较大的域差异。在驾驶数据集上训练的模型通常在监控场景中只能得到较低质量的结果。此外，对于从 SYNTHIA 到 Cityscapes 的无监督域自适应（UDA），HRDA [38] 达到了 65.8% 的 mIoU 分数。然而，HRDA 在监控场景中仅取得了 45.4% 的 mIoU 分数，这也表明了驾驶场景和监控场景之间的巨大域差异。提供一个高质量的人工标注数据集用于分析监控场景中不同方法的有效性，将对社区有很大帮助。这使研究人员能够对交通监控场景中的分割方法、实例分割方法和无监督分割方法的有效性进行交叉验证。

4. 分割方法评估

4.1. 实现细节

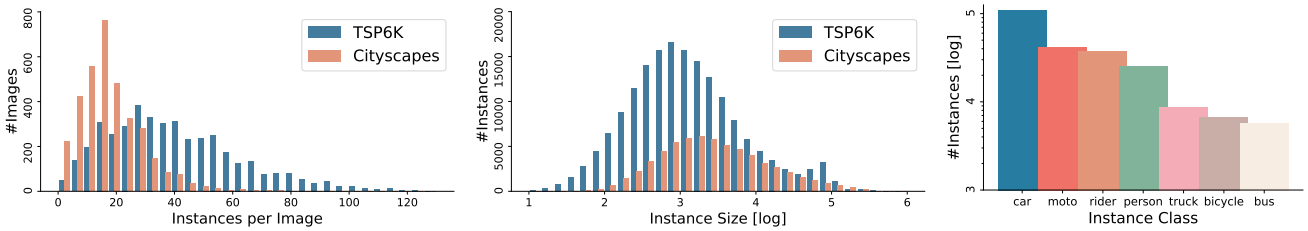
我们在一个流行的代码库 mmsegmentation [16] 上运行所有场景解析方法。所有模型均在配备 8 块 NVIDIA A100 GPU 的节点上训练。对于基于 CNN 的方法，输入尺寸设置为 769×769. 对于基于 Transformer 的方法，输入尺寸设置为 1024×1024. 我们对所有方法进行 160,000 次迭代训练，批量大小为 16。

由于 SETR [100] 占用大量 GPU 内存，因此输入尺寸设置为 769×769，批量大小设置为 8。此外，我们使用 mmsegmentation [16] 中的默认数据增强方法。我们使用 mIoU [57] 指标来评估场景解析方法的性能。正如 [17] 中提到的，mIoU 指标对大尺寸的对象实例存在偏倚。然而，监控交通场景中充满了小型交通参与者。为了更好地评估交通参与者的实例，我们参照 [17] 的方法，使用 iIoU 指标，覆盖所有包含实例的类别。

4.2. 性能分析

不同方法的评估结果见表 3。场景解析方法可以大致分为几个组。接下来，我们主要讨论使用编码器-解码器结构、自注意力机制和变换器结构的方法。

编码器-解码器结构： 基于编码器-解码器结构的方法利用高分辨率的低层特征来细化分割图的细节。UperNet [80] 和 DeepLabv3+ [12] 将编码器-解码器结构应用于分割网络。与 DeepLabv3 [11] 相比，DeepLabv3+ [12] 利用高分辨率特征，能进一步提升分割结果，在两个数据集上分别提高了超过 0.6% 的 mIoU 分数和 1% 的 iIoU 分数。我们观察到，编码



(a) (b) (c)

图 3. TSP6K 数据集的数据分析：(a) 每张图像中实例数量的分布。(b) 实例尺寸的分布。(c) 每个类别的实例数量。

表 3. 以前的场景解析方法在 TSP6K 验证集和测试集上的评估结果。

Methods	Publication	Backbone	Parameters	GFlops	Validation		Test	
					mIoU (%)	iIoU (%)	mIoU (%)	iIoU (%)
FCN [57]	CVPR'15	R50	49.5M	454.1	71.5	55.2	72.5	55.1
PSPNet [98]	CVPR'16	R50	49.0M	409.8	71.7	54.8	72.6	54.8
DeepLabv3 [11]	ArXiv'17	R50	68.1M	619.3	72.4	55.0	73.3	55.0
UperNet [80]	ECCV'18	R50	66.4M	541.0	72.4	55.2	73.1	55.0
DeepLabv3+ [12]	ECCV'18	R50	43.6M	404.8	73.1	56.1	73.9	56.3
PSANet [99]	ECCV'18	R50	59.1M	459.2	71.3	54.5	72.6	54.8
EMANet [48]	ICCV'19	R50	42.1M	386.8	72.0	55.5	72.9	55.5
EncNet [92]	CVPR'18	R50	35.9M	323.3	71.4	54.8	72.7	55.0
DANet [24]	CVPR'19	R50	49.9M	457.3	72.3	56.0	73.1	56.1
CCNet [41]	ICCV'19	R50	49.8M	460.2	72.0	55.3	73.1	55.3
KNet-UperNet [96]	NeurIPS'21	R50	62.2M	417.4	72.6	56.8	73.7	56.5
OCRNet [88]	ECCV'20	HR-w18	12.1M	215.3	73.2	55.3	73.7	55.1
SETR [100]	CVPR'21	ViT-Large	310.7M	478.3	70.5	44.9	70.7	45.0
SegFormer [83]	NeurIPS'21	MIT-B2	24.7M	72.0	72.9	54.6	73.8	54.9
SegFormer [83]	NeurIPS'21	MIT-B5	82.0M	120.8	74.5	56.7	74.8	56.7
Swin-UperNet [56]	ICCV'21	Swin-Base	121.3M	1184.6	74.9	57.4	75.6	57.2
SegNeXt [27]	NeurIPS'22	MSCAN-Base	27.6M	80.2	74.6	57.3	75.4	57.2
SegNeXt [27]	NeurIPS'22	MSCAN-Large	48.9M	258.6	74.8	57.7	75.6	57.6
DRD (Ours)	–	**MSCAN-Base**	**46.1M**	**90.1**	**75.8**	**58.4**	**75.9**	**58.0**

器-解码器结构对小型物体分割非常有用，其中 iIoU 分数有了显著提升。

自注意力机制在场景解析方法中被广泛使用在评估的方法中，EncNet [92]、DANet [24]、EMANet [48] 和 CCNet [41] 都采用了不同类型的自注意力机制。它们中的大多数比 FCN 表现更为出色。由于自注意力机制能够建模长距离的像素依赖关系，这些方法能够细化最终的分割结果，从而提升性能。我们观察到，EncNet 相比于 FCN 没有性能提升。我们分析认为，EncNet 利用通道级自注意力机制来构建全局上下文，但这种机制无法很好地保留局部细节，尤其是在包含不同大小交通参与者的交通监控场景中。

Transformer 结构已成功应用于计算机视觉任务 [6, 20]，通常能比卷积神经网络结构获得更好的识别结

表 4. 以前的实例分割方法在 TSP6K 验证集和测试集上的评估结果。过程中我们基于 ResNet-50 [32] 主干网络运行所有方法。

Methods	Validation		Test	
	AP_{box}	AP_{seg}	AP_{box}	AP_{seg}
YOLACT [4]	19.9	13.7	20.7	14.6
Mask-RCNN [31]	27.2	23.5	27.0	23.5
SOLO [75]	–	29.6	–	29.8
SOLOv2 [76]	–	28.6	–	28.6
QueryInst [23]	37.7	31.5	37.2	31.3
Mask2Former [14]	32.9	31.3	32.5	31.4

果。SETR [100]、Segformer [83]、Swin-UperNet [56] 和 SegNeXt [27] 都将 Transformer 作为场景解析的骨干网络。其中，SETR 的解析结果明显较差，而其他变换器结构的表现优于卷积骨干网络。其中，使用 Swin [56] 骨干网络的 UperNet [80] 在 mIoU 和 iIoU 指标上都比使用 ResNet-50 [32] 骨干网络的表现要好得多。此外，与 Swin-UperNet 相比，SegNeXt 在性能上相似，但仅使用了大约 20% 的参数和 7% 的 GFlops。

总之，编码器-解码器结构、空间自注意力机制和 Transformer 结构是提升交通监控场景解析的有效策略。在第 7节中，我们根据这些策略设计了一个更强大的解码器，可以进一步改进 SegNeXt，其表现优于 Hamburger 解码器 [26]。

5. 实例分割方法评估

我们为 TSP6K 中的每张交通图像提供了额外的实例标注，这些标注可用于评估实例分割方法在分割和分类交通监控图像中的交通参与者（即人和车辆）方面的性能。交通参与者的类别如下：行人、骑车人、汽车、卡车、公交车、摩托车和自行车。我们评估了几种经典的实例分割方法，包括 YOLACT [4]、Mask RCNN [31]、SOLO [75]、SOLOv2 [76]、QueryInst [23] 和 Mask2Former [14]。上述所有方法都基于公开的代码库 mmdetection [8] 进行实验。平均精度（AP）指标的结果见表 4。

性能分析： 在评估的方法中，QueryInst [23] 的表现优于其他方法。它还取得了最佳的 7.4% AP_s 分数。这个较差的性能表明现有方法在小型实例分割上存在困难。此外，我们观察到基于 ResNet-50 的 Mask-RCNN 在 Cityscapes 上取得了 40.9% 的框 AP 和 36.4% 的掩码 AP，这些结果比在 TSP6K 上训练的模型高出超过 10% 的 AP。性能差异表明，在 TSP6K 上的实例分割仍然是一个巨大的挑战。我们希望额外的实例标注能帮助社区提升实例分割方法在监控场景中分割交通参与者的性能。

表 5. 无监督领域适应方法的评估结果

Methods	SYNTHIA → TSP6K		Cityscapes → TSP6K	
	mIoU (%)	Imprv (%)	mIoU (%)	Imprv (%)
Baseline	21.7	0	26.1	0
ADVENT [73]	22.3	+0.6	31.7	+5.6
DA-SAC [2]	33.0	+11.3	33.9	+7.8
SePiCo [81]	33.8	+12.1	35.9	+9.8
DAFormer [37]	33.4	+11.7	39.5	+13.4
HRDA [38]	45.4	+23.7	54.1	+18.0

6. 无监督域自适应

近年来，场景解析的无监督域自适应（UDA）方法得到了广泛研究。然而，大多数 UDA 方法集中于将合成驾驶场景迁移到真实驾驶场景中。得益于提出的 TSP6K 数据集，我们可以研究将驾驶场景迁移到交通监控场景的 UDA 方法。具体而言，我们进行了 UDA 实验，将 SYNTHIA [63] 和 Cityscapes [17] 数据集分别迁移到 TSP6K 数据集。我们选择了几种经典的 UDA 方法进行评估，包括 ADVENT [73]、DA-SAC [2]、SePiCo [81]、DAFormer [37] 和 HRDA [38]。实验结果见表 5。请注意，我们仅计算和平均了源域和目标域中共同类别的结果。

性能分析： 我们建立了一个基线，该基线在源域上训练 DeepLab-v2 [10]，并直接在目标域上进行推断。所有评估的 UDA 方法都显著超越了基线。我们可以观察到，近期基于 Transformer 的 UDA 方法在性能上优于基于 CNN 的 UDA 方法。UDA 方法的最佳性能仍然远逊于完全监督的方法（54.1% 对 72.4%）。此外，从 Cityscapes 到 TSP6K 的 UDA 性能明显高于从 SYNTHIA 到 TSP6K 的 UDA。这表明现有的驾驶数据集可以促进对交通监控场景的理解。我们希望提出的数据集能够促进 UDA 方法在交通监控场景解析任

务中的发展。

7. 提出的场景解析方法

如 第 3节中分析的，交通监控场景通常捕捉到比驾驶场景更多的交通内容，而且不同语义区域的尺度和形状差异更大。此外，小物体和背景物体占据了很大比例。这些情况使得准确解析这些场景变得具有挑战性。为了适应交通监控场景，我们提出了一种细节优化解码器。我们解码器的设计原则有两个方面。

第一点，由于骨干网络最后的特征空间分辨率非常低，基于低分辨率特征构建解码器通常会生成粗糙的解析结果，从而大大影响小物体的分割。正如一些先前的研究所验证的 [53, 68, 75]，低级高分辨率特征对分割小物体非常有帮助。因此，我们利用编码器-解码器结构来融合低分辨率和高分辨率特征，以改善小物体的分割。**第二点**，如在 第 4节中分析的，自注意力机制是编码空间信息进行场景解析的有效方法。然而，直接将自注意力机制应用于高分辨率特征会消耗大量计算资源，特别是在处理高分辨率交通场景图像时。受到 [14] 中为每个区域学习表示的启发，我们提出引入多个区域标记，并在每个区域标记与来自高分辨率特征的每个 patch 标记之间建立成对的关联。

7.1. 整体流程

我们基于 第 4节中总结的宝贵经验构建了适用于交通监控场景的场景解析网络。首先，我们采用了 SegNeXt [27] 中提出的强大编码器作为我们的编码器，该编码器在我们的 TSP6K 数据集上以低计算成本实现了良好的结果。然后，我们在该编码器上构建了一个细节优化解码器（DRD）。细节优化解码器的流程如图 4所示，包含两个部分。对于第一部分，我们遵循 DeepLabv3+ [12] 的解码器设计来生成细粒度的特征图。其中 ASPP 模块直接添加到编码器中。请注意，我们没有使用来自第二阶段的 $\times 4$ 下采样特征，而是使用了来自第三阶段的 $\times 8$ 特征，如 [27] 中建议的那样。第二部分是区域细化模块，其细节将在下一个小节中描述。

7.2. 区域细化模块

区域细化模块旨在精细化交通图像中的不同语义区域。形式上，设 $\mathbf{F} \in \mathbb{R}^{HW \times C}$ 为解码器第一部分的展平特征，其中 H、W 和 C 分别表示特征图的高度、宽度和通道数。设 $\mathbf{R} \in \mathbb{R}^{N \times C}$ 为 N 个可学习的区域标记，每个标记是一个 C 维向量。展平特征 $\mathbf{F}$ 和可学习的区域标记 $\mathbf{R}$ 分别通过三个线性层生成查询（query）、键（key）和值（value），具体如下：

$$\mathbf{R}_Q, \mathbf{F}_K, \mathbf{F}_V = f_Q(\mathbf{R}), f_K(\mathbf{F}), f_V(\mathbf{F}), \tag{1}$$

其中，$f_Q(\mathbf{R})$、$f_K(\mathbf{F})$ 和 $f_V(\mathbf{F})$ 是线性层，$\mathbf{R}_Q \in \mathbb{R}^{N \times C}$、$\mathbf{F}_K \in \mathbb{R}^{HW \times C}$ 和 $\mathbf{F}_V \in \mathbb{R}^{HW \times C}$ 分别是生成的查询、键和值。我们计算 $\mathbf{F}$ 和 $\mathbf{R}$ 之间的多头交叉注意力，具体如下：

$$\mathbf{R}_E = \text{Softmax}\left(\frac{\mathbf{R}_Q \mathbf{F}_K^T}{\sqrt{C}}\right)\mathbf{F}_V + \mathbf{R}, \tag{2}$$

其中，$\mathbf{R}_E \in \mathbb{R}^{N \times C}$ 是得到的区域嵌入。然后，区域嵌入被送入一个前馈网络，其公式为：

$$\mathbf{R}_O = \text{FFN}(\mathbf{R}_E) + \mathbf{R}_E, \tag{3}$$

其中，$\mathbf{R}_O$ 是前馈网络的输出。在这里，按照 [69] 的做法，仅将区域标记 $\mathbf{R}_E$ 送入前馈块以提高处理效率。

接下来，将 $\mathbf{R}_O$ 和 $\mathbf{F}$ 传递到两个线性层，以生成一组新的查询和键，具体如下：

$$\mathbf{R}_{Q1}, \mathbf{F}_{K1} = f_{Q1}(\mathbf{R}_O), f_{K1}(\mathbf{F}). \tag{4}$$

我们对 $\mathbf{R}_{Q1}$ 和 $\mathbf{F}_{K1}$ 进行矩阵乘法，以生成注意力图，计算如下：

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{R}_{Q1} \mathbf{F}_{K1}^T}{\sqrt{C}}\right), \tag{5}$$

其中，$\mathbf{A} \in \mathbb{R}^{N \times HW}$ 表示 N 个注意力图，每个注意力图对应一个语义区域。当我们获得区域注意力图后，我们通过广播乘法将 $\mathbf{A}$ 和 $\mathbf{F} \in \mathbb{R}^{HW \times C}$ 结合起来，具体如下：

$$\mathbf{S}_{i,j,k} = \mathbf{A}_{i,j} \cdot \mathbf{F}_{j,k}, \tag{6}$$

其中，$\mathbf{S} \in \mathbb{R}^{N \times HW \times C}$ 是输出。最后，$\mathbf{S}$ 被重新排列并调整形状为 $\mathbb{R}^{N \times C \times H \times W}$，然后送入一个卷积层以生成最终的分割图。

8. 实验

表 3 列出了不同方法的性能表现。可以看出，我们的方法超越了所有以前的方法，在两个指标上都取

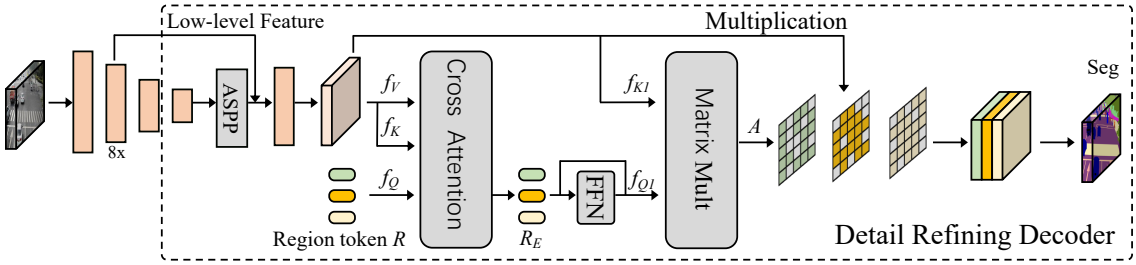


图 4. 细节细化解码器的流程。我们的解码器分为两部分。第一部分类似于 DeeplabV3+ [12] 中提出的解码器。不同的是，我们使用来自第三阶段（与输入相比下采样 ×8）的特征图来融合 ASPP 中的特征图。第二部分是提出的区域细化模块。

得了最佳结果。为了验证提出的细节精化解码器的有效性，我们进行了若干关于区域词元数量和注意力头的消融实验。由于篇幅限制，实验细节和消融结果放在了附录材料中。

9. 结论

在本文中，我们构建了 TSP6K 数据集，专注于交通监控场景。我们为每张交通图像提供了语义和实例标签。基于精细标注的 TSP6K 数据集，我们还评估了一些流行的场景解析方法、实例分割方法和 UDA 方法。了提高场景解析的性能，我们设计了一个细节优化解码器，该解码器利用来自编码-解码结构的高分辨率特征，并基于区域精化模块对不同的语义区域进行精化。细节精化解码器学习了多个区域标记，并计算了不同语义区域的注意力图。注意力图用于精化不同语义区域中的像素相似度。实验结果表明细节优化解码器是有效的。

方法的局限性： 该数据集包含 6,000 张标注图像。我们还有大量未标注的图像，可以进一步探索。数据集在地理上不够多样化，缺乏来自左侧行驶国家的场景。TSP6K 的数据仅包含 RGB 图像，这限制了多模态模型的发展。

致谢： This work was in part supported by National Key R&D Program of China (No. 2022ZD0118700), NSFC (NO. 62276145), the Fundamental Research Funds for the Central Universities (Nankai University, 070-63223049), CAST through Young Elite Scientist Sponsorship Program (No. YESS20210377). Computations were supported by the Supercomputing Center of Nankai University (NKSC).

References

[1] Aau rainsnow traffic surveillance dataset, 2018. 4

[2] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15384–15394, 2021. 7

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017. 3

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Int. Conf. Comput. Vis.*, pages 9157–9166, 2019. 3, 7

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1209–1218, 2018. 1

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229. Springer, 2020. 6

[7] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4974–4983, 2019. 3

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Seman-

tic image segmentation with deep convolutional nets and fully connected crfs. In *Int. Conf. Learn. Represent.*, 2015. 3

[10] 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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017. 3, 7

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5, 6

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.*, pages 801–818, 2018. 3, 5, 6, 8, 9

[13] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S Huang, Wen-Mei Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In *Int. Conf. Comput. Vis.*, pages 5218–5228, 2019. 1, 3

[14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1290–1299, 2022. 3, 7, 8

[15] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inform. Process. Syst.*, 34:17864–17875, 2021. 3

[16] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 5

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3213–3223, 2016. 1, 2, 3, 4, 5, 7

[18] Jonathan Crespo, Jose Carlos Castillo, Oscar Martinez Mozos, and Ramon Barber. Semantic information for robot navigation: A survey. *Applied Sciences*, 10(2):497, 2020. 1

[19] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *Proc. Int. Conf. Intelligent Transportation Systems*, pages 3819–3824, 2018. 2

[20] 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 *Int. Conf. Learn. Represent.*, 2021. 3, 6

[21] Boyang Du, Congju Du, and Li Yu. Megf-net: multi-exposure generation and fusion network for vehicle detection under dim light conditions. *Visual Intelligence 1*, 2023. 1

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

[23] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Int. Conf. Comput. Vis.*, pages 6910–6919, 2021. 3, 7

[24] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3146–3154, 2019. 6

[25] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 4

[26] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? In *Int. Conf. Learn. Represent.*, 2021. 7

[27] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *Adv. Neural Inform. Process. Syst.*, 2022. 2, 3, 6, 7, 8

[28] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational visual media*, 2023. 3

[29] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022. 3

[30] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7519–7528, 2019. 1

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 3, 7

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 7

[33] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Int. Conf. Mach. Learn.*, pages 1989–1998. Pmlr, 2018. 3

[34] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1335–1344, 2018. 3

[35] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *arXiv preprint arXiv:2211.11943*, 2022. 3

[36] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4003–4012, 2020. 3

[37] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9924–9935, 2022. 7

[38] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 372–391. Springer, 2022. 5, 7

[39] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2702–2719, 2019. 2

[40] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6409–6418, 2019. 3

[41] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 603–612, 2019. 3, 6

[42] Galadrielle Humblot-Renaux, Letizia Marchegiani, Thomas B Moeslund, and Rikke Gade. Navigation-oriented scene understanding for robotic autonomy: Learning to segment driveability in egocentric images. *IEEE Robotics and Automation Letters*, 7(2):2913–2920, 2022. 1

[43] Rui Jiang, Ruixiang Zhu, Hu Su, Yinlin Li, Yuan Xie, and Wei Zou. Deep learning-based moving object segmentation: Recent progress and research prospects. *Machine Intelligence Research*, 20(3):335–369, 2023. 1

[44] Xiaojie Jin, Huaxin Xiao, Xiaohui Shen, Jimei Yang, Zhe Lin, Yunpeng Chen, Zequn Jie, Jiashi Feng, and Shuicheng Yan. Predicting scene parsing and motion dynamics in the future. In *Adv. Neural Inform. Process. Syst.*, volume 30, 2017. 1

[45] Jean-Philippe Jodoin, Guillaume-Alexandre Bilodeau, and Nicolas Saunier. Urban tracker: Multiple object tracking in urban mixed traffic. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 885–892. IEEE, 2014. 4, 5

[46] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[47] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11593–11603, 2022. 3

[48] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-

maximization attention networks for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 9167–9176, 2019. 6

[49] Zhetao Li, Ziwen Chen, Wei-Shi Zheng, Sangyoon Oh, and Kien Nguyen. Ar-cnn: an attention ranking network for learning urban perception. *Science China Information Sciences*, 65(1):112104, 2022. 1

[50] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Int. Conf. Comput. Vis.*, pages 6758–6767, 2019. 3

[51] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 752–761, 2018. 1

[52] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1925–1934, 2017. 1, 3

[53] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2117–2125, 2017. 8

[54] 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 *Eur. Conf. Comput. Vis.*, 2014. 2

[55] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8759–8768, 2018. 3

[56] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 6, 7

[57] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, 2015. 1, 3, 5, 6

[58] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2014. 1

[59] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12435–12445, 2021. 3

[60] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Int. Conf. Comput. Vis.*, pages 4990–4999, 2017. 2, 4

[61] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Eur. Conf. Comput. Vis.*, pages 102–118. Springer, 2016. 3

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[63] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3234–3243, 2016. 3, 7

[64] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Int. Conf. Comput. Vis.*, pages 7374–7383, 2019. 2

[65] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Int. Conf. Comput. Vis.*, pages 10765–10775, 2021. 2, 4

[66] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 7262–7272, 2021. 1, 3

[67] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang.

Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8797–8806, 2019. 4, 5

[68] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Eur. Conf. Comput. Vis.*, pages 282–298. Springer, 2020. 3, 8

[69] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Int. Conf. Comput. Vis.*, pages 32–42, 2021. 8

[70] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7472–7481, 2018. 3

[71] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Int. Conf. Comput. Vis.*, pages 1456–1465, 2019. 3

[72] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 1743–1751. IEEE, 2019. 2

[73] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2517–2526, 2019. 3, 7

[74] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 642–659. Springer, 2020. 3

[75] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *Eur. Conf. Comput. Vis.*, pages 649–665. Springer, 2020. 3, 7, 8

[76] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Adv. Neural Inform. Process. Syst.*, 33:17721–17732, 2020. 3, 7

[77] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 9092–9101, 2021. 3

[78] Dong Wu, Man-Wen Liao, Wei-Tian Zhang, Xing-Gang Wang, Xiang Bai, Wen-Qing Cheng, and Wen-Yu Liu. Yolop: You only look once for panoptic driving perception. *Machine Intelligence Research*, 19(6):550–562, 2022. 1

[79] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yu Yizhou. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. In *arXiv preprint arXiv:1903.11816*, 2019. 3

[80] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Eur. Conf. Comput. Vis.*, pages 418–434, 2018. 5, 6, 7

[81] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 7

[82] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12193–12202, 2020. 3

[83] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inform. Process. Syst.*, 34:12077–12090, 2021. 1, 3, 6, 7

[84] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4085–4095, 2020. 3

[85] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.*, pages 1–18, 2021. 3

[86] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 325–341,

2018. 3

[87] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2636–2645, 2020. 1, 2, 4

[88] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Eur. Conf. Comput. Vis.*, 2020. 6

[89] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context for semantic segmentation. *Int. J. Comput. Vis.*, 129(8):2375–2398, 2021. 3

[90] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash-creating hazard-aware benchmarks. In *Eur. Conf. Comput. Vis.*, pages 402–416, 2018. 2, 4

[91] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying panoptic segmentation for autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21351–21360, 2022. 2

[92] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7151–7160, 2018. 6

[93] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Adv. Neural Inform. Process. Syst.*, 32, 2019. 3

[94] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Int. Conf. Comput. Vis.*, pages 2031–2039, 2017. 1

[95] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12083–12093, 2022. 3

[96] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *Adv. Neural Inform. Process. Syst.*, volume 34, pages 10326–10338, 2021. 6

[97] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Eur. Conf. Comput. Vis.*, pages 405–420, 2018. 3

[98] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2881–2890, 2017. 1, 3, 6

[99] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Eur. Conf. Comput. Vis.*, pages 267–283, 2018. 3, 6

[100] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6881–6890, 2021. 1, 3, 5, 6, 7

[101] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 633–641, 2017. 1, 2

[102] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 593–602, 2019. 1

[103] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Eur. Conf. Comput. Vis.*, pages 289–305, 2018. 3

[104] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Int. Conf. Comput. Vis.*, pages 5982–5991, 2019. 3