arXiv:2303.01991v2 [cs.CV] 2 Apr 2023

统一感知：以最低标注成本实现高效的深度感知视频全景分割

Kurt Stolle[1]　　Gijs Dubbelman[1]

摘要——深度感知视频全景分割是实现基于摄像机的场景理解的一种很有前景的方法。然而，与基于图像的方法相比，目前最先进的方法需要成本高昂的视频标注，并采用复杂的训练流程。本文提出了一种名为“统一感知”的新方法，无需基于视频的训练即可达到最先进水平。我们的方法采用简单的两阶段级联跟踪算法，复用基于图像的网络计算得到的目标嵌入。在 Cityscapes-DVPS 数据集上的实验结果表明，该方法的总体 DVPQ 达到 57.1，超过了最先进的方法。此外，我们还证明，我们的跟踪策略能够有效地在 KITTI-STEP 上进行长期目标关联，STQ 达到 59.1，超过了采用相同骨干网络的最先进方法。

一、引言

本文提出了一种面向智能车辆的统一移动感知方法，通过单一的端到端网络，将全景分割、单目深度估计和目标跟踪结合起来。我们的目标是开发一个能够实时、有效且高效地执行这些任务的框架，使车辆能够做出明智的决策，并提升整体安全性和性能。同时，我们旨在减少训练网络所需的标注工作量，这是视频感知系统开发中的主要瓶颈。

近年来，深度学习的发展显著提升了深度感知视频全景分割等联合计算机视觉任务的性能 [1]–[4]。然而，这些方法通常依赖于在完整标注的视频序列上训练的跟踪模块，这限制了其向其他可能缺乏此类标注的模态的泛化能力 [5]。此外，与基于图像的方法相比，训练这些方法需要多得多的标注工作，因此难以对大规模数据集进行标注 [6]。通过使用一个跟踪模块，复用在基于图像的深度感知全景分割网络中作为中间数据而已计算出的目标嵌入，我们无需进行基于视频的训练，即可达到当前最先进的性能。图1展示了这一概念的概览。

我们的方法基于最先进的全景分割网络，并对其进行扩展以估计每个实例的深度图。由于跟踪器使用基于图像的推理过程中生成的中间目标嵌入，

* 本研究由 NEON项目 资助，该项目部分经费来自 荷兰研究委员会（NWO）。[1] Kurt Stolle 和 Gijs Dubbelman 任职于荷兰埃因霍温理工大学移动感知系统（MPS）实验室。

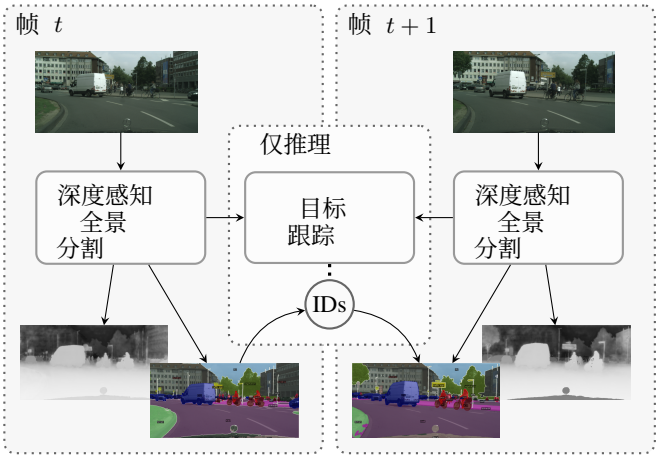


图 1：我们提出的统一感知框架概览。训练仅使用图像完成，降低了流程的复杂性，并减少了相较于基于视频的方法所需的标注工作量。推理时，网络中已计算出的中间数据会被（重新）用于关联不同帧中的对象。

因此无需额外的视频训练流程。我们在 Cityscapes-DVPS [2] 数据集上通过实验验证了我们的方法，其准确率和标注工作量均优于最先进的方法。此外，我们在 KITTI-STEP [5], 上展示了我们方法的长期跟踪能力，并表明每帧生成的中间嵌入具有足够的代表性，能够持续跟踪目标。我们还分析了跟踪模块增加的推理时间，发现其可以忽略不计，因此我们的方法适用于实时应用。

这项工作的主要贡献是一种新的深度感知视频全景分割方法，无需基于视频的训练和高成本标注工作，即可达到最先进水平。我们的方法采用一个简单的跟踪器，复用基于图像的网络计算出的目标嵌入，证明这些中间数据足以在整个视频中表示物体，并可用于全景跟踪。所提出的方法为视频场景理解提供了一种更高效、更实用的解决方案，在自动驾驶、机器人技术和监控领域具有潜在应用。我们的代码可在 tue-mps.github.io/unipercept 获取。

II. 相关工作

A. 全景分割

全景分割 [7] 是一项旨在准确分割图像中的物体实例和背景区域的任务，从而全面理解场景。该任务的一般方法包括采用两阶段流程 [8]–[11], ，分别执行实例分割和语义分割，然后将两者融合。另一种方法是采用单个网络 [12]–[15] ，同时预测实例级和语义级分割图。Panoptic FCN 方法[12]是一种单阶段方法，已证明借助一种优雅的动态卷积方法能够达到最先进水平；该方法的特点是，每个检测到的物体都有一个经过学习的压缩表示，即 卷积核，可用于跟踪。

B. 单目深度估计

单目深度估计的任务是根据单幅图像估计场景的深度图。历史上，该任务通常使用手工设计特征和几何约束来解决 [16]–[18]。近年来，基于深度学习的方法得到广泛应用，许多研究表明，可以训练网络从单幅图像中预测具有高准确率的深度图 [19]–[21]。此外，近期研究还探索了在单目深度估计中使用分割信息 [3], [4], [22], ，其中通过以每个掩码处的实例级特征为条件来预测深度估计值。全景深度 [4]便是此类方法，它利用实例级分割掩码对每个物体的深度估计值进行归一化。

近年来，自监督单目深度估计（SSMDE）作为一种解决缺乏无噪声深度真值标注问题的有前景方法，受到了广泛关注 [22]–[26]。该方法利用视频中连续帧之间的自然几何约束来训练单目深度估计模型。Collaborative Competition [27] 使用深度平滑损失进行深度感知的全景分割，该损失基于估计深度图的梯度，并由输入图像的像素强度梯度加权。我们的方法将这种深度平滑损失与实例级深度归一化 [4]相结合，以改善深度的物体级表示，从而为目标跟踪提供稳健的空间线索。

C. 目标跟踪

目标跟踪是指在视频序列中按时间关联目标检测结果，通常会用到空间和外观线索 [28]–[31]。在全景分割中，这意味着为视频中出现的每个独立 事物实例分配一致且唯一的标识符。近期有多种方法在视频全景分割领域探索了这一问题 [32], [33]。

物体跟踪数据的标注既耗费成本又耗时，因此近期研究着重开发无监督或弱监督物体跟踪方法 [34], [35]。在视频全景分割中，标注成本尤其高昂，因为除了要持续识别每个目标外，还必须仔细绘制并标注其全景分割掩码 [6]。我们提出了一种深度感知视频全景分割网络，通过利用用于分割和深度估计的中间物体表示来执行目标跟踪，从而免去了这项标注工作。这使我们能够以基于图像的方式训练网络，相较于现有需要完整视频序列真值标注的方法，这是一个显著优势。

III. 方法

在本节中，我们介绍名为“统一感知”的框架，该框架整合了全景分割、跟踪和单目深度估计任务。为此，我们扩展了一种最先进的统一网络架构，该架构采用动态卷积进行全景分割和单目深度估计。图 2 展示了这一用于基于图像的全景分割和单目深度估计的架构。此外，通过利用该网络在推理过程中生成的中间数据，我们提出了一种基于关联的跟踪算法，利用每一帧生成的外观和空间线索，为视频序列中的每个物体有效分配唯一身份。

A. 全景分割

我们以 Panoptic FCN [12] 框架作为网络的基础。该架构提出使用动态卷积进行单路径全景分割。具体而言，该网络由骨干网络、特征生成器、位置头和掩码核生成器组成。特征生成器为图像中的每个像素生成一个嵌入$\mathbf{E}_m$ ，掩码核生成器则为位置头估计的每个实例中心位置$\mathbf{L}$ 提供一个掩码核$\mathbf{K}_m$ 。随后，将掩码核与嵌入相乘，即执行动态卷积，从而为每个实例生成一个分割掩码$\mathbf{M}$。

该方法的一个特性是，每个检测到的物体在$\mathbf{K}_m$中都有一个对应的核向量，该向量经过训练后能够嵌入实例级视觉特征。因此，对于场景中的每个物体，这个卷积核都是其外观的唯一压缩表示。此外，我们假设，同一物体在两个连续视频帧中的核向量具有一定的相似度，足以将其与其他物体区分开来，从而用于跟踪。正如稍后将在第III-D节中详细说明的那样，我们利用这一特性在推理时跨帧关联物体，而无需进行基于视频的训练。

B. 单目深度估计

如图 2 所示，我们为网络扩展了用于单目深度估计的深度核和特征生成器分支

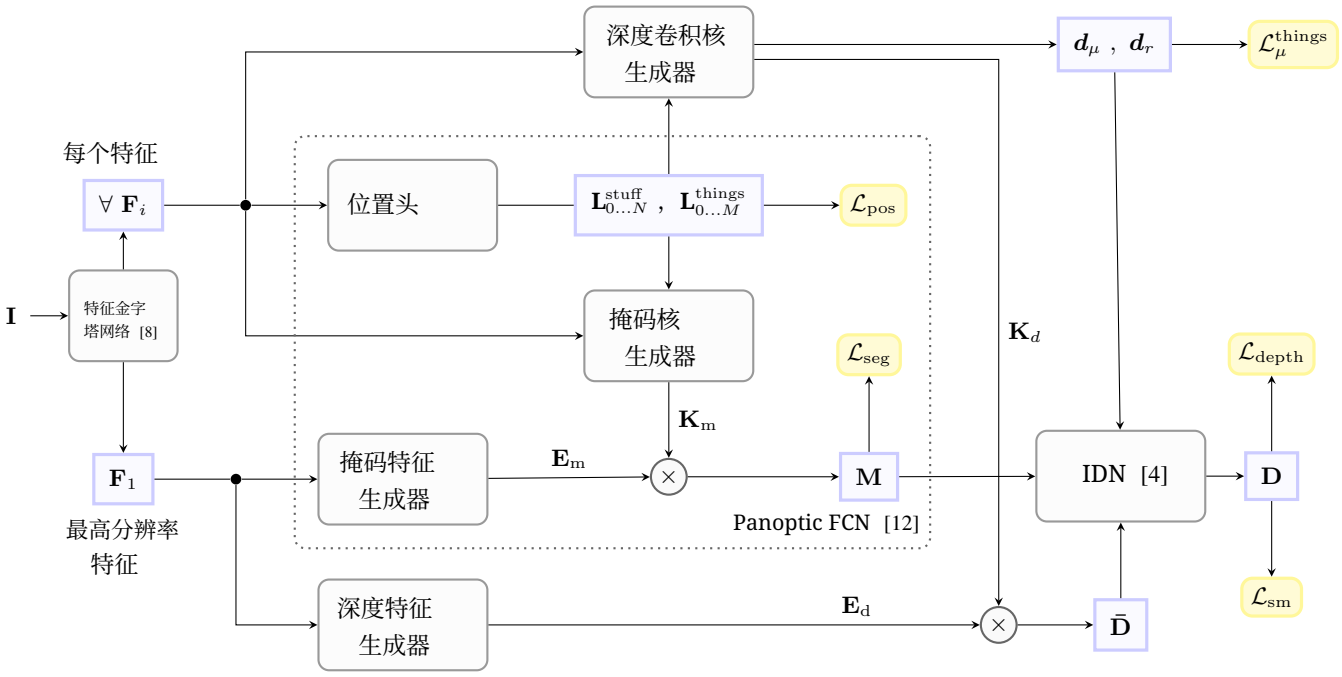


图 2：我们的全景分割和深度估计架构。

。遵循动态卷积范式，对于每个检测到的实例，都会生成深度核嵌入，并在与掩码核嵌入相同的位置 $\mathbf{L}$ 进行采样。随后，该深度核 $\mathbf{K}_d$ 与深度嵌入$\mathbf{E}_d$ 相乘，再经过 Sigmoid 激活，得到归一化深度图

$$\bar{\mathbf{D}} = \sigma(\mathbf{K}_d \cdot \mathbf{E}_d) \in [-1, 1] \tag{1}$$

针对每个实例，这在流程上与我们在第三节-A中生成实例掩码的方式相似。随后，我们实施 Panoptic Depth [4] 中提出的逐实例深度归一化流程。这包括将每个归一化深度图的值偏移平均深度$\boldsymbol{d}_\mu$ ，然后乘以范围 $\boldsymbol{d}_r$；这两个量均针对每个实例进行估计。正如稍后在第 III-E 节中所述，这些反归一化变量提供了关于每个实例位置的对象级信息，可作为跟踪的有用线索。按照惯例，我们将尺度不变对数误差与相对平方误差结合起来，作为深度估计任务的损失函数 [2], [4], [36]。

此外，我们注意到，真实深度图通常含有噪声和生成伪影。正如在 Mono DVPS [3], 中通过实验验证的，自监督平滑损失可以提高深度估计器对这类真实值噪声的鲁棒性。基于 [27], 中提出的方法，我们为每个¯实例级深度图$\mathbf{D}$ 定义像素强度加权平滑损失，形式如下

$$\mathcal{L}_{\text{sm}} = \left|\partial_x \bar{\mathbf{D}}\right| e^{-|\partial_x \mathbf{I}|} + \left|\partial_y \bar{\mathbf{D}}\right| e^{-|\partial_y \mathbf{I}|} \tag{2}$$

其中， x 和 y 分别表示水平方向和垂直方向，$\mathbf{I}$ 表示转换为灰度的输入图像。

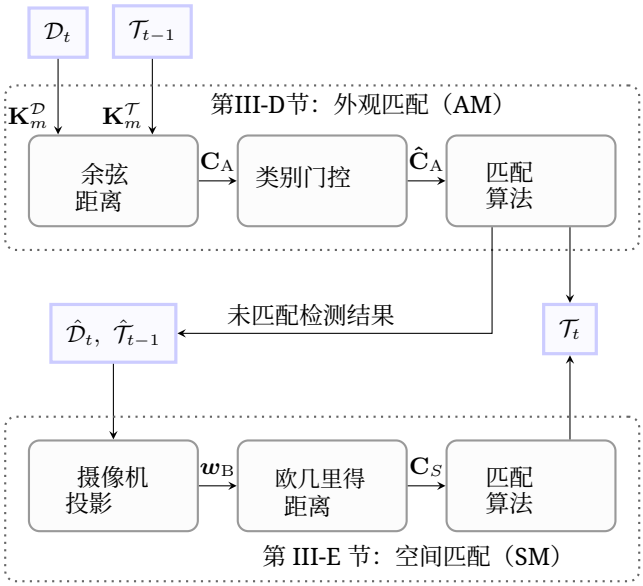


图 3：我们提出的跟踪架构采用两阶段分配算法，利用外观和空间线索，其中 $\mathcal{D}$ 是检测结果集， $\mathcal{T}$ 是轨迹片段集。

C. 目标跟踪

在每一帧中，跟踪器 模块都需要为视频序列中出现的每个独特事物 实例分配一个时间上保持一致的轨迹标识符。跟踪器会应用于之后的每一帧（在线），因此无法获取未来帧的信息。其结构

用于定义网络在某一帧提出的检测 $\delta \in \mathcal{D}$ 的，是一个由网络中间数据组成的元组，即

$$\delta = (\mathbf{L}, \mathbf{K}_m, \boldsymbol{d}_\mu). \tag{3}$$

当跟踪器将标识符 $\iota \in \mathbb{N}$ 分配给某个检测结果时，由此形成的结构通常称为轨迹片段 $\tau \in \mathcal{T}$，其表示为

$$\tau = (\delta, \iota) \tag{4}$$

我们将检测结果依次输入多个基于跟踪的关联阶段。每个阶段都会根据线性和分配的形式进行匹配。具体而言，我们计算一个代价函数

$$\mathbf{C} : (\mathcal{D}, \mathcal{T}) \longrightarrow \mathbb{R}_+^{|\mathcal{D}|\times|\mathcal{T}|} \tag{5}$$

该函数表示当前时间步 t 的所有检测结果 $\mathcal{D}_t$与上一帧中的轨迹片段 $\mathcal{T}_{t-1}$ 之间的分配代价。接着，我们建立一个线性分配问题，以得出匹配

$$\mathcal{M} \subset \{(\delta_i, \tau_n) \cdots (\delta_j, \tau_m) \mid i \neq j, n \neq m\} \tag{6}$$

使得

$$\mathcal{M}^\star = \arg\min_{\mathcal{M}} \sum_{m \in \mathcal{M}} \mathbf{C}(m) \tag{7}$$

其中， $\mathcal{M}^\star$ 表示检测结果与轨迹片段的最优组合，可使总分配代价最小。我们使用 Jonker-Volgenant算法 [37] 求解这一问题。

每个匹配轨迹片段的标识符 ι 都会传递给相应的检测结果，而未匹配的检测结果和轨迹片段则会传递至下一个跟踪器阶段。最后，所有剩余的未匹配检测结果都会被分配新生成的标识符。

图 3 展示了我们提出的跟踪器示意图。通过采用两个级联阶段，分别基于掩码核向量$\mathbf{K}_m$中的外观线索（第 III-D 节）和实例平均深度 $\mathbf{d}_\mu$ 中的空间线索（第 III-E 节）定义代价函数，我们旨在构建一个跟踪流程，使其能够抵御因基于视频的训练未能确保每个被重复使用的目标表示保持一致而产生的噪声。相比之下，单阶段跟踪仅采用外观表示或空间表示。

D. 外观匹配（*AM*）阶段

外观匹配阶段能够依据目标的视觉特征，有效地关联连续帧中的目标。在我们的方案中，掩码核向量$\mathbf{K}_m$是对目标视觉特征的紧凑表示。我们根据检测结果 $\mathcal{D}$ 和轨迹片段 $\mathcal{T}$的掩码核向量，通过余弦相似度定义代价函数；余弦相似度用于计算两个向量之间的夹角。因此，代价函数定义为

$$\mathbf{C}_\mathrm{A}(\mathbf{K}_m^\mathcal{D}, \mathbf{K}_m^\mathcal{T}) = 1 - \frac{\mathbf{K}_m^\mathcal{D} \cdot \mathbf{K}_m^\mathcal{T}}{||\mathbf{K}_m^\mathcal{D}|| \; ||\mathbf{K}_m^\mathcal{T}||} \tag{8}$$

其中，将相似度从 1 中减去以得到一个最小化目标。按照惯例， [34],我们对代价函数应用类别门控，将预测类别标签不同的条目设为具有无限代价。

E. 空间匹配（*SM*）阶段

鉴于实例平均深度 $\boldsymbol{d}_\mu$ 包含对每个目标平均深度的估计，我们旨在定义一个利用这项中间数据作为代理的代价函数，以衡量目标从前一帧到当前帧的位置相似度。为此，我们对摄像机和待跟踪目标的运动作出两个假设。首先，我们假设摄像机安装在阿克曼驱动系统上，这是一种车辆中常用的转向机构。这意味着摄像机无法进行纯旋转。其次，我们假设目标主要在水平面上移动，这一假设对于大多数智能车辆应用都成立。基于这些假设，我们提出在场景中创建目标的二维俯视图表示，将目标坐标映射到与地面平行的水平面上的点。我们使用相机内参、平均深度 $\boldsymbol{d}_\mu$以及图像$\mathbf{L}$中的目标中心坐标，将每个目标投影到该平面上。我们采用欧氏距离作为空间代价函数，它表示场景中两个位置之间的相似度，即

$$\mathbf{C}_\mathrm{S}(\boldsymbol{w}_\mathrm{B}^\mathcal{D}, \boldsymbol{w}_\mathrm{B}^\mathcal{T}) = ||\boldsymbol{w}_\mathrm{B}^\mathcal{D} - \boldsymbol{w}_\mathrm{B}^\mathcal{T}|| \tag{9}$$

其中 $\boldsymbol{w}_\mathrm{B}$ 表示鸟瞰视图中的物体坐标。

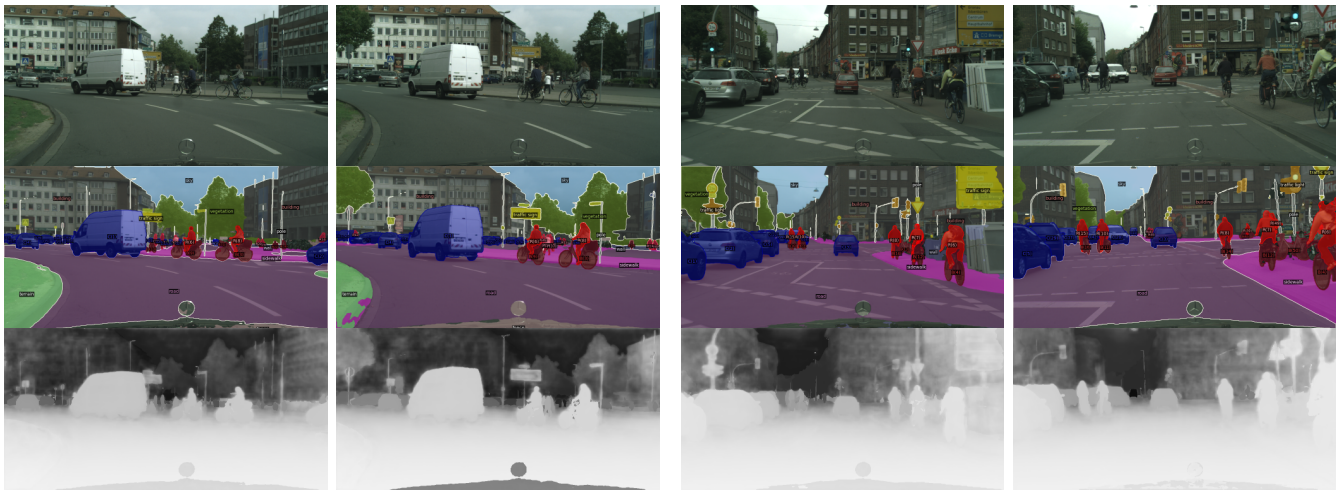
IV. 实验

在本节中，我们通过实验评估所提出方法的性能。首先，我们在 Cityscapes-DVPS [2]上执行深度感知全景分割任务。这展示了网络在考虑所有目标任务时的整体性能：全景分割、单目深度估计和目标跟踪。其次，我们重点关注在 KITTI-STEP 数据集 [5]上进行的全景跟踪。在这里，我们通过消融所提出的跟踪阶段，研究网络的长期关联能力。

A. 数据集

Cityscapes-DVPS： 我们采用 ViP-DeepLab [2], 提出的 Cityscapes-DVPS 基准，该基准扩展自 Cityscapes-VPS [32], 并添加了通过立体视差计算得到的深度图。该数据集中的每个序列包含 30 帧，其中每第 5 帧都标注了全景分割和深度图。按照标准评估流程，结果使用深度感知视频全景质量（DVPQ）指标进行解读。该指标本质上衡量的是仅考虑绝对相对深度误差低于阈值的像素时的视频全景质量 [32] 。

KITTI-STEP： 我们旨在使用 KITTI-STEP 数据集 [5], 评估我们的方法在长期跟踪方面的能力，该数据集通过分割和



(a) 明斯特 28　　(b) 明斯特 78

图4：Cityscapes-DVPS 数据集上的定性结果。

表I：在消融深度归一化和平滑损失的情况下，Cityscapes-DVPS 数据集上的结果

方法	实例深度归一化 a	深度平滑损失 b	骨干网络	视频训练必需	DVPQ ↑	$DVPQ^{th}$ ↑	$DVPQ^{st}$ ↑
统一感知（我们的方法）	No	No	ResNet-50	No	43.9	33.2	50.5
统一感知（我们的方法）	Yes	No	ResNet-50	No	50.4	44.3	53.4
统一感知（我们的方法）	No	Yes	ResNet-50	No	45.4	36.0	51.4
统一感知（我们的方法）	Yes	Yes	ResNet-50	No	**57.1**	**57.4**	57.6
PolyphonicFormer [36]			ResNet-50	Yes	48.1	35.6	57.1
Mono DVPS [3]			ResNet-50	Yes	50.4	35.9	**61.7**
ViP-DeepLab [2]			WR-41	Yes	55.1	43.3	**63.6**

[a]启用实例深度归一化 [4]，否则所有深度图都会归一化到相同尺度。b 在所有训练和微调过程中启用边缘感知深度平滑损失 [27]。

跟踪质量（STQ）进行评估。该指标定义为关联质量（AQ）与分割质量（SQ）的几何平均值，其中关联质量用于衡量跟踪性能。将关联得分与分割得分分开评估非常重要，因为这样我们便可以独立于全景分割性能评估我们方法的跟踪性能。该数据集采用 Cityscapes 标注方案，因此训练可以从 Cityscapes-DVPS 模型的训练权重开始。此外，由于 KITTI-STEP 数据集不提供真实深度图，我们仅使用深度平滑损失（第 III-B 节）来训练单目深度估计模块。虽然这对于深度估计任务而言并不理想，但我们发现，它足以生成对跟踪有用的对象级深度估计。

B. 实现细节

我们使用 ResNet-50 [38] 作为骨干网络，并在 COCO [39]上对其权重进行预训练。随后，我们按照 Panoptic FCN [12]的相同训练设置，在 Cityscapes [40] 上训练网络 80K 步。真值深度图由数据集补充材料中提供的立体视差图生成。我们注意到，这些深度图很可能是通过与用于生成 Cityscapes-DVPS 50 [38] 中所提供深度图的专有算法不同的方法计算得出的，这导致两个数据集之间存在轻微的域差距。然而，我们通过经验发现，加入深度平滑损失（第 III-B 节）似乎能够有效地正则化网络，使其学习到一种泛化表示，不会因模态变化而受到负面影响。接下来，我们使用 AdamW 优化器和 1×10^{-8}的学习率，在从 Cityscapes-DVPS [2]序列中打乱抽取的帧集合上对模型进行微调，训练 20K 步。最后，对于 KITTI-STEP [5] 数据集，我们使用相同的超参数额外微调 10K 步。

C. 主要结果

在表I中，我们将提出的框架与 Cityscapes-DVPS [2] 上的最先进水平进行比较。在我们的解读中，我们仅与使用相同骨干网络训练的模型进行比较，因为我们发现使用更大的骨干网络会导致比较有失公平。与所评估的方法相比，我们的方法取得了显著的性能提升。这主要归功于 *thing* 类得分的提高；我们认为这得益于 Panoptic Depth [4] 提出的实例深度归一化。此外，实验

表 II：KITTI-STEP 数据集上跟踪器模块消融实验的结果

方法	跟踪器阶段[a]	骨干网络	视频训练必需	STQ ↑	AQ ↑	SQ ↑
统一感知（我们的方法）	AM	ResNet-50	No	56.8	51.9	61.9
统一感知（我们的方法）	SM	ResNet-50	No	53.8	46.9	61.9
统一感知（我们的方法）	AM→SM	ResNet-50	No	**59.1**	**56.4**	**61.9**
Motion-DeepLab [5]		ResNet-50	Yes	52.2	45.6	59.8
SIAin [41]		Swin-L	Yes	57.9	55.2	60.7
TubeFormer-DeepLab [42]		Axial-ResNet-50	Yes	**65.2**	**61.0**	**70.3**

[a]我们跟踪算法中启用的阶段，见图 3

表 III：图形处理器上的推理时间 T （毫秒）

	Cityscapes-DVPS		KITTI-STEP	
	T	Δ [b]	T	Δ [b]
基线 a	154	–	70	–
AM	160	6	73	3
SM	164	10	76	6
AM→SM	165	11	77	7

[a]已禁用所有跟踪模块。对象标识符随机分配。b 与基线的差异。

结果表明，单独使用深度平滑损失 [27] 无法取得有竞争力的性能。然而，尽管要达到有竞争力的性能并非必须使用该损失，但将其与深度归一化结合使用，可使我们模型的性能显著优于现有方法。图4展示了定性可视化结果。

KITTI-STEP [5]上的结果汇总于表 II，并针对各跟踪器阶段进行了消融实验。我们发现，AM 和 SM 阶段组成的完整级联明显优于基线网络 Motion-DeepLab [5]，其 STQ 达到 59.1。此外，尽管由于训练时缺少真实值，我们的深度图可能并非最优，但平均深度在跟踪中具有重要价值：仅使用 SM 阶段便取得了 46.9 的 AQ 分数，超过基线 AQ 分数 45.6。这可能是因为平均深度能够很好地近似与摄像机之间的距离，而距离是跟踪中的一个关键因素。

最后，我们通过分析各跟踪阶段对推理时间的影响，探讨了本方法在部署方面的表现。我们在单块 NVIDIA RTX A6000 GPU 上以批量大小为 1，对 10 万帧的平均推理时间进行了性能分析。作为基线，我们分析了未使用跟踪算法时网络的运行情况。需要指出的是，这种配置无法产生有效结果，但可以计算各阶段带来的额外开销。表 III 总结了本实验的结果。在 Cityscapes-DVPS 和 KITTI-STEP 上，我们完整的两阶段跟踪器分别为全景分割模型和单目深度估计模型的推理时间增加了 11 毫秒和 7 毫秒。考虑到本方法带来的显著性能提升，这一开销是可以接受的。此外，我们尚未针对生产环境对模型进行优化。我们相信，这类工作将进一步缩短推理时间，使现代基于 GPU 的嵌入式系统能够实现实时推理。

V. 结论

我们提出了统一感知（Unified Perception），这是一个联合执行全景分割、单目深度估计和目标跟踪的框架。由于我们的模型仅使用图像进行训练，与那些在视频序列上学习目标表示的最先进替代方法相比，部署该方法所需的标注成本大幅降低。

在我们的方法中，我们假设，基于图像的深度感知全景分割网络所生成的中间物体表示，包含了足以促进跟踪的信息。实验结果确实表明，与其他基于视频的方法相比，我们提出的方法在 Cityscapes-DVPS 和 KITTI-STEP 上取得了显著更好的性能。我们指出，采用规模更大、表达能力更强的骨干网络（如 Swin Transformer [43] 或 Axial-ResNet [44],）有望进一步提升性能，这将留待未来工作探索。

尽管其他深度感知视频全景分割方法为场景理解的新颖建模方式提供了宝贵见解，但我们的方法表明，要取得有竞争力的结果，并不一定需要通过基于视频的训练来增加模型复杂度。我们希望这项研究能启发人们以新的方式，利用（再利用）中间对象描述符来处理各种任务，从而降低标注成本。

参考文献

[1] X. Li, W. Zhang, J. Pang, K. Chen, G. Cheng, Y. Tong, and C. C. Loy, 《Video K-Net：一种简单、强大且统一的视频分割基线》，载于*CVPR*，2022年。[2] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, 《ViP-DeepLab：通过深度感知视频全景分割学习视觉感知》，载于*2021*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*。IEEE。[3] A. Petrovai and S. Nedevschi, 《MonoDVPS：一种用于深度感知视频全景分割的自监督单目深度估计方法》，第3076–3085页。[4] N. Gao, F. He, J. Jia, Y. Shan, H. Zhang, X. Zhao, and K. Huang, *PanopticDepth*：深度感知全景分割的统一框架。arXiv。[5] M. Weber, J. Xie, M. Collins, and Y. Zhu, 《STEP：分割并跟踪每个像素》，研究技术报告。

[6] P. Meletis, X. Wen, C. Lu, D. de Geus, and G. Dubbelman, 《用于场景理解的Cityscapes-Panoptic-Parts和PASCAL-Panoptic-Parts数据集》，arXiv，技术报告，zSCC: 0000009 arXiv:2004.07944 [cs, eess] 类型：文章。 [在线]。可访问：http://arxiv.org/abs/ 2004 07944 。A. Kirillov, K. He, R. B. Girshick, C. Rother, and P. Dollar, 《全景分割》，第9396–9405页。[8] A. Kirillov, R. Girshick, K. He, and P. Dollar, 《全景特征金字塔网络》，*IEEE*计算机学会计算机视觉与模式识别会议论文集，第2019年6月卷，第6392–6401页，2019年1月。[9] H. Liu, C. Peng, C. Yu, J. Wang, X. Liu, G. Yu, and W. Jiang, 《用于全景分割的端到端网络》，第6165–6174页。[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, 《Deeplab：结合深度卷积网络、空洞卷积和全连接条件随机场的语义图像分割》，第40卷，第834–848页。[11] J. Lazarow, K. Lee, and Z. Tu, 《学习全景分割中的实例遮挡》，第10 717–10 726页。[12] Y. Li, H. Zhao, X. Qi, L. Wang, Z. Li, J. Sun, and J. Jia, 《用于全景分割的全卷积网络》。[13] K. Sofiiuk, O. Barinova, a nd A. Konushin, 《Adaptis：自适应实例选择网络》，第7354–7362页。[14] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, 《Panoptic-DeepLab：一种简单、强大且快速的自底向上全景分割基线》，载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*。IEEE。[15] L.-C. Chen, H. Wang, and S. Qiao, 《面向全景分割的宽残差网络扩展》，第abs/2011.11675卷。[16] D. Hoiem, A. A. Efros, and M. Hebert, 《从单幅图像获取几何上下文》，第1卷，第654–661页，第1卷。[17] A. Saxena, M. Sun, and A. Y. Ng, 《Make3d：从单幅静态图像学习三维场景结构》，第31卷，第5期，第824–840页。[18] M. Liu, M. Salzmann, and X. He, 《从单幅图像进行离散-连续深度估计》，第716–723页。[19] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, 《利用全卷积残差网络进行更深层次的深度预测》，第239–248页。[20] J. Li, R. Klein, and A. Yao, 《一种从单幅RGB图像估计精细深度图的双流网络》，第3392–3400页。[21] A. Atapour-Abarghouei and T. Breckon, 《利用合成数据和通过图像风格迁移实现的域自适应进行实时单目深度估计》，第2800–2810页。[22] M. Klingner, J.-A. Term¨ohlen, J. Mikolajczyk, and T. Fingscheidt, 《自监督单目深度估计：通过语义引导解决动态物体问题》，第abs/2007.06936卷。[23] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, 《从视频中无监督学习深度和自我运动》，载于*2017*年*IEEE*计算机视觉与模式识别会议 *(CVPR)*，第6612–6619页。[24] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman, 《通过观察静止的人来学习运动中人物的深度》，第4516–4525页。[25] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, 《深入探究自监督单目深度估计》，载于*IEEE/CVF*国际计算机视觉会议论文集 *(ICCV)*。[26] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, 《结合结构和语义进行无监督单目深度与自我运动学习》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*）研讨会。[27] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, 《竞争性协作：深度、相机运动、光流和运动分割的联合无监督学习》，载于*2019*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第12 232–12 241页。[28] Z. Wang, H. Zhao, Y.-L. Li, S. Wang, P. H. S. Torr, and L. Bertinetto, 《不同的跟踪任务是否需要不同的外观模型？》[29] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, 《Fairmot：论多目标跟踪中检测与再识别的公平性》，第129卷，第3069–3087页。[30] B. Yan, Y. Jiang, P. Sun, D. Wang, Z. Yuan, P. Luo, and H. Lu, 《迈向目标跟踪的大统一》。[31] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, 《用于深度行人再识别的强基线与批量归一化颈部》，第22卷，第10期，第2597–2609页。

[32] D. Kim、S. Woo、J.-Y. Lee 和 I. S. Kweon, 《视频全景分割》。[33] W. Wang、T. Zhou、F. Porikli、D. Crandall 和 L. Van Gool, 《视频分割深度学习技术综述》。[34] X. Q. Zhang、R. H. Jiang、C. X. Fan、T. Y. Tong、T. Wang 和 P. C. Huang, 《视觉跟踪深度学习方法进展：文献综述与基础》，《国际自动化与计算期刊》*2021 18:3*，第18卷，第311–333页，2021年3月。[35] F. Bastani、S. He 和 S. Madden, 《具有跨输入一致性的自监督多目标跟踪》。[36] H. Yuan、X. Li、Y. Yang、G. Cheng、J. Zhang、Y. Tong、L. Zhang 和 D. Tao, 《Polyphonicformer：面向深度感知视频全景分割的统一查询学习》。[37] R. Jonker 和 A. Volgenant, 《一种用于稠密和稀疏线性分配问题的最短增广路径算法》，第38卷，第4期，第325–340页。[38] K. He、X. Zhang、S. Ren 和 J. Sun, 《用于图像识别的深度残差学习》，第770–778页。[39] T.-Y. Lin、M. Maire、S. Belongie、L. Bourdev、R. Girshick、J. Hays、P. Perona、D. Ramanan、C. L. Zitnick 和 P. Doll´ar, 《Microsoft COCO：上下文中的常见物体》，arXiv，技术报告。[40] M. Cordts、M. Omran、S. Ramos、T. Rehfeld、M. Enzweiler、R. Benenson、U. Franke、S. Roth 和 B. Schiele, 《用于城市场景语义理解的Cityscapes数据集》，发表于*2016*年*IEEE*计算机视觉与模式识别会议（*CVPR*）。IEEE。[41] J. Ryu 和 K. Yoon, 《一种使用Transformer的端到端可训练视频全景分割方法》，第abs/2110.04009卷。[42] D. Kim、J. X. H. W. S. Qiao、Q. Yu 和 H.-S. Kim, 《Tubeformer-Deeplab：视频掩码Transformer》。[43] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo, 《Swin Transformer：使用移位窗口的分层视觉Transformer》，第9992–10 002页。[44] H. Wang、Y. Zhu、B. Green、H. Adam、A. Yuille 和 L.-C. Chen, 《Axial-DeepLab：用于全景分割的独立轴向注意力》，arXiv，技术报告，zSCC: 0000443 2003 07853 arXiv：。[ ] [ ] 2003 07853 cs类型：文章。在线。可访问：http://arxiv.org/abs/。