基于多任务 transformer 的智能驾驶感知框架

张邦梅

（贵州大学 省部共建公共大数据国家重点实验室，贵州 贵阳 550025）

摘要：视觉智能感知是自动驾驶领域中的一项关键任务，其通过实时监控驾驶道路环境来实现车辆安全行驶。单独的驾驶感知任务性能已经达到瓶颈，开发新的技术以提升任务性能是一个艰难的挑战。当前多任务协同的智能感知是一个有效的解决思路，其利用任务间的有用信息来提升所学任务泛化性能。为了应对这个挑战，我们以驾驶感知中两个关键任务（车道线检测和驾驶道路区域分割）为例，提出一种多任务学习框架，其利用这两个任务间的相关性来提升检测任务和分割任务性能。首先，通过骨干网络提取输入图像的特征。然后，通过 transformer 提取图像场景的全局特征，并且在编码器和解码器中分别为驾驶道路区域分割任务和车道线检测任务设置不同的检测头。在此基础上，道路分割检测头通过 U-net 网络实现道路区域分割，车道线检测头通过多层感知机实现出车道线路径划分。最后，通过自动加权求和多个损失函数来同时学习多个任务。我们在 BerkeleyDeepDrive100K（BDD100K）数据集上验证该框架的有效性。实验结果表明，该框架在各个指标中均显著优于当前流行的多任务和单任务方法，并保持每秒超过 36.4 帧的实时推理。

关键词：多任务；车道线检测；道路区域分割；视觉 transformer

由于环境的高度复杂性，驾驶感知是智能驾驶系统最具挑战性的任务之一。在驾驶感知任务中，驾驶道路区域分割和车道线检测是其中两个关键任务。此外，为了限制车辆的机动，视觉感知系统应该能够理解场景，然后为决策系统提供信息，包括：道路是否可通行的判断、车道线的位置等。然而，目前大多数驾驶感知任务都集中在单个任务处理。因此，多任务协同的全景驾驶感知系统是未来智能驾驶感知系统的必然趋势，对可行驶区域分割和车道检测，以帮助车辆遵守交通规则。在以往的研究中，视觉感知任务是单独运行处理的。例如，车道线检测框架 SCNN 和 ENet-SAD；语义分割方法 PSPNet 和 SegNet。这些研究在各自领域取得了显著成果，但智能驾驶视觉感知需要同时处理多个任务，以协同运行的方式进行视觉感知更有利于汽车的安全智能驾驶。最近，基于深度学习的多任务学习方法可以协同地处理多个相关的任务，通过任务间的参数共享来提高的预测性能。

在这项研究中，我们提出了一种高效的端到端多任务学习框架来加速驾驶感知。框架由一个 transformer 编码器 – 解码器和两个检测头组成。在编码器维度和解码器维度分别设置检测头用于驾驶道路区域分割和车道线检测，两个感知任务在编码器中实现参数共享。我们工作的主要贡献可以总结如下：1. 我们设计了一个多任务学习框架，可以同时处理驾驶感知的两个关键任务。2. 我们在编码维度和解码维度加入两个检测头分别用于道路区域分割和车道线检测。3. 与对比的多任务和单任务模型相比，所提出的方法在 BDD100K 数据集上实现优异的性能。

一、相关工作

（一）车道线检测

车道线分割是驾驶感知的基础任务，已经进行了广泛的尝试来划分道路上的可行驶车道。Pan 等人提出了一种用于交通场景理解的空间卷积神经网络（SCNN），相邻像素之间的消息传递有效地保持了车道线结构的连续性。ENet-SAD 通过在网络自身执行自上而下和分层的注意力蒸馏网络进行进一步表征学习，允许一个模型从自身学习并获得实质性的改进，而不需要任何额外的监督或标签。

（二）可行驶的区域分割

深度学习的巨大成功出现在语义分割领域。全卷积网络（FCN）尝试使用可端到端训练的深度学习执行语义分割，他们使用 1x1 卷积和上采样进行分割。PSPNet 使用金字塔场景解析网络提取各种缩放特征。尽管在基于深度学习的分割任务中实现了显著的准确性，但推理时间仍然是一个遗留问题。SegNet 提出了一种新颖的解码器网络，它将低分辨率编码器特征图上采样为全输入分辨率特征图。少量可训练参数提供了良好的性能和具有竞争力的推理时间。

二、算法实现

（一）vision transformer

视觉 transformer 的具体网络模型如图 1 所示：

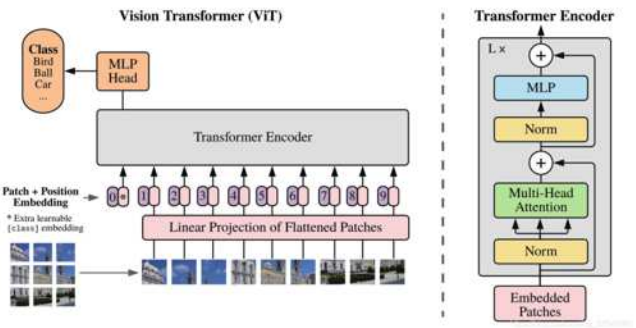

图 1　视觉 transformer 框架

1.transformer 编码器

如图 1 右所示，transformer 编码器首先要经过一个网络层均一化（Layer Norm）处理，在进入多头注意力机制（Multi-Head Attention）层前通过同一个特征变换操作生成了 Q、K、V 三个向量，多头注意力机制通过计算 Q，K 向量内积来评价图像特征间的关联性，然后将获得 softmax 权值与 V 向量进行内积生成原特征向量感兴趣区域的特征映射。多头自注意力机制的公式定义如下：

$$\text{Attention}(Q, K, V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

其中 d_k 作为归一化系数把内积值重新缩放回均值为 0，方差为 1 的状态。最后通过由全连接层 +*GELU* 激活函数 +*Dropout* 组成多层感知机层对编码器输出特征进行处理。

2.transformer 解码器

常规的视觉 transformer 只采用了 transformer 的编码器部分，通过多层感知机检测头对编码器的特征输出进行计算，从而输出标签类别。在本论文中，我们加入由 DETR 六个编码器层组成标准的 Transformer 解码器之上。进一步在高维特征中提取全局信息，以加强车道线检测的准确度，框架图如 2 所示。

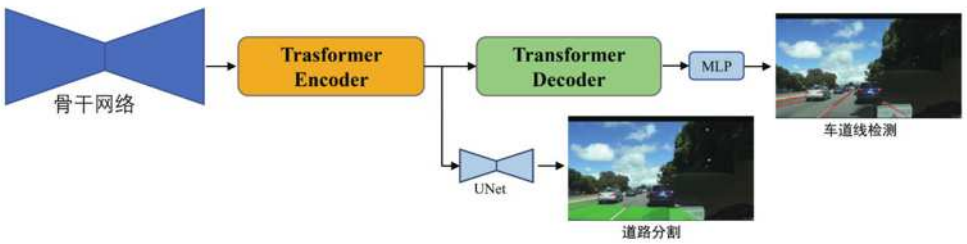

图 2 智能驾驶感知框架

（二）车道线检测与道路识别

1. 车道线检测

车道线检测头通过多层感知机（MLP）推断出车道线路径划分，MLP 层将输出特征维度转化为车道线标签特征维度，如公式（2）所示

$$f'=\text{Pro}\ (f\in\mathbb{R}^{\times h\times w}) \quad (2)$$

其中 Proj 表示由 CNN 和 MLP 组成的神经元传播方式，输出特征为网络模型定义的超参，本文使用 256 的输出特征维度。处理后的特征维度与车道线标签维度统一，最后将 MLP 层输出结果与车道线标签进行二分类的交叉熵损失函数：

$$L=\frac{1}{N}\sum_i -[y_i\log(p_i)+(1-y_i)\log(1-p_i)] \quad (3)$$

2. 驾驶道路识别分割

在用于驾驶道路识别分割的 U-net 模块中，每一个卷积块多个卷积层、ReLU 激活函数及批量归一化构成，将不同尺度的特征图通过特征变化到相同维度。解码器部分由几个解码阶段组成。其中每次特征变换的空间维度变化分别为 128、256、256 和 512，并且每一次特征融合后相同尺度的特征具有相同的特征通道。

三、实验

（一）数据集

我们通过 BDD100K 数据集上的最新方法进行比较来验证所提出方法的有效性。BDD100K 数据集已发布用于自动驾驶研究。它包括可驾驶区域、对象检测、属性、道路类型和车道标记的各种注释。 帧的一些特定属性，例如天气、场景和一天中的时间。天气状况的类型包括下雨、下雪、晴天、阴天、部分多云、有雾和不确定。一天中的时间包括白天、夜晚、黎明 / 黄昏和未定义。在此数据集上进行训练时，环境和天气的多样性提高了我们网络的鲁棒性。BDD100K 数据集由 1，280，720 张图像组成。总共 100K 的图像被分成三个分割；70K 用于训练，10K 用于验证，20K 用于测试。

（二）实验环境

我们使用深度学习框架 Pytorch 实现整体网络模型，使用 Adam 作为优化器进行训练，学习率 1×10-4，初始学习率在每 15 个 epoch 降低一半。我们在 BDD100K 数据集训练网络，在 NVIDIA Tesla V00 GPU 上训练了 200 个 epoch，Batchsize 设置为 24。

（三）评价指标

我们使用以下指标与竞争方法进行评估和定量比较：可行驶区域分割的 mIoU（%）；车道线分割的准确率（%）和 IoU（%），我们还使用每秒帧数（fps）比较了算法处理速度。在评价预测结果与真实样本间的关系时，TP 表示实际为正预测为正的情况，TN 表示实际为负预测为负情况，FP 表示实际为负预测为正的情况，FN 表示实际为正预测为负的情况。其中 IoU 表示预测区域与真实区域相交面积与合并面积的比值，mIoU 为交并比的均值：

$$IoU=\frac{TP}{TP+FP+FN} \quad (4)$$

$$mIoU=\frac{sum(IoU)}{class} \quad (5)$$

其中 class 表示预测的类别数。车道线检测准确度（Accuracy）

表示预测像素类别正确的像素数占总像素数的比例：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

（四）实验结果

与对比方法（表 1）相比，所提出的网络显示了最先进的可驱动区域分割性能。 所提出的方法达到了 91.68% 的 mIoU，高于所有对比方法。并且我们所提出的方法可以以 36.4fps 的速度实时执行。这对于安全至关重要且资源有限的驾驶辅助应用来说是一个显著优势。表 2 将我们的车道线分割结果与其他最先进的方法进行了比较。所提出的方法与其他车道线分割方法相比具有最高的性能，分别具有 71.23% 和 24.86% 的准确率和 IoU 分数。图 3 显示了所提出的可驾驶区域分割和车道线检测结果，可以看到车道线检测分割效果较好，并且在无明显道路线环境下能够正常识别驾驶道路区域。

表 1 BDD100K 数据集上的可行驶区域分割结果。

Methods	mIoU（%）	Speed（fps）
MultiNet[14]	71.60	8.6
DLT-Net[15]	72.10	9.3
PSPNet[5]	89.60	11.1
Ours	91.68	36.4

表 2 BDD100K 数据集上的车道线检测结果。

Methods	Accuracy（%）	IoU（%）
ENet[16]	34.12	14.64
ENet-SAD[4]	36.56	16.02
SCNN[3]	35.79	15.84
Ours	71.23	24.86

图 3　车道线检测和道路区域结果示例

四、结论

本文提出了一种多任务的智能驾驶感知框架。在这项研究中，我们提出了一种多任务学习框架，其有效地利用驾驶区域分割任务和车道线检测任务间的有用信息来提升各自任务的泛化性能。首先，通过骨干网络提取输入图像高分辨率特征。然后，通过 transformer 提取图像场景特征全局信息，并且在编码器维度和解码器维度分别设置检测头用于驾驶道路区域分割和车道线检测。最后，通过多个损失函数的加权求和来学习不同的任务。实验证明，我们的算法取得了优异的效果，并且能够实时检测车道线以及分割驾驶道路区域。

参考文献：

[1] 李轩，王飞跃 . 面向智能驾驶的平行视觉感知：基本概念、框架与应用 [J]. 中国图象图形学报，2021，26（01）：67-81.

[2] 王世峰，戴祥，徐宁，等 . 无人驾驶汽车环境感知技术综述 [J]. 长春理工大学学报（自然科学版），2017，40（01）：1-6.

[3] 叶伟，朱明 . 基于空间特征聚合的车道线检测算法 [J]. 计算机系统应用，2021，30（12）：235-242.

[4] 刘彬，刘宏哲 . 基于改进 Enet 网络的车道线检测算法 [J]. 计算机科学，2020，47（04）：142-149.

[5] 祁欣，袁非牛，史劲亭，王贵黔 . 多层次特征融合网络的语义分割算法 [J]. 计算机科学与探索，2022，1-13.

[6] 梅迪 . 应用于图像语义分割的神经网络——从 SegNet 到 U-Net[J]. 电子制作，2021（12）：49-52.

[7] 王汉谱，瞿玉勇，刘志豪，等 . 基于 FCN 的图像语义分割算法研究 [J]. 成都工业学院学报，2022，25（01）：36-41.

[8] 田永林，王雨桐，王建功，等 . 视觉 Transformer 研究的关键问题：现状及展望 [J]. 自动化学报，2022，48（04）：957-79.

[9] 刘文婷，卢新明 . 基于计算机视觉的 Transformer 研究进展 [J]. 计算机工程与应用，2022，58（06）：1-16.

[10] 柳胜超 . 复杂背景下交通信号灯检测与识别方法研究与应用 [D]. 长安大学，2021.

[11] 唐闻 . 基于深度学习的计算机图像识别技术研究 [J]. 电脑编程技巧与维护，2022（01）：154-156+166.

[12] 李明，来国红，常晏鸣，等 . 深度学习算法中不同优化器的性能分析 [J]. 信息技术与信息化，2022（03）：206-209.

[13] 孙宇菲 . 基于多任务学习的车道线检测算法研究 [D]. 长安大学，2021.

本文系 2022 年贵州省基础研究计划（自然科学项目）“基于茶树树叶病害图像的细粒度识别研究”（黔科合基础 -ZK[2022] 一般 108）。