Computer Science and Application 计算机科学与应用, 2024, 14(4), 141-150
Published Online April 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.144085

Hans汉斯

基于残差U块和上下文变换器的三分支实时语义分割

冉照彬，王　超

北京信息科技大学计算机学院，北京

收稿日期：2024年3月8日；录用日期：2024年4月10日；发布日期：2024年4月18日

摘　要

针对有限的卷积接受域阻碍了全局关系建模的问题，本文提出一种基于残差U块和上下文变换器的三分支实时语义分割算法，该网络采用空间信息、上下文信息、边界信息三个并行的分支结构，并且采用不同深度的残差U块构建网络的上下文信息分支来获取更具鲁棒性的多尺度特征。同时增加上下文变换器模块来增强全局关系建模能力。通过实验表明了该方法的有效性，在Cityscapes数据集上，没有使用预训练的情况下可以在单个V100上使用全分辨率图像(1024 × 2048)以76.5 FPS的速度达到78.6% MIoU。

关键词

实时语义分割，全局关系建模，多尺度特征，上下文变换器

Trilateral Network with Residual U-Blocks and Contextual Transformer Block for Real-Time Semantic Segmentation

Zhaobin Ran, Chao Wang

School of Computing, Beijing Information Science and Technology University, Beijing

Received: Mar. 8th, 2024; accepted: Apr. 10th, 2024; published: Apr. 18th, 2024

Abstract

To address the problem where the limited receptive field of convolutions hinders the modeling of global relationships, this paper proposes a three-branch real-time semantic segmentation algorithm based on residual U-blocks and context transformers. The network employs three parallel

文章引用: 冉照彬, 王超. 基于残差 U 块和上下文变换器的三分支实时语义分割[J]. 计算机科学与应用, 2024, 14(4): 141-150. DOI: 10.12677/csa.2024.144085

branch structures for spatial information, contextual information, and boundary information, utilizing residual U-blocks of varying depths to build the network's contextual information branch to obtain more robust multi-scale features. Additionally, a context transformer module is introduced to enhance the capability for global relationship modeling. Experiments demonstrate the effectiveness of this method; on the Cityscapes dataset, without the use of pretraining, it can achieve 78.6% MIoU at a speed of 76.8 FPS on a single V100 GPU using full-resolution images (1024 × 2048).

Keywords

Real-Time Semantic Segmentation, Global Relationship Modeling, Multi-Scale Features, Context Transformer Module

1. 引言

语义分割是计算机视觉中的一项重要任务，它旨在将输入图像的每个像素划分到特定的类别中。随着深度学习的发展，语义分割已经成为众多领域的关键技术，包括自动驾驶[1]、医学影像诊断[2]以及遥感成像[3]等。自从全卷积网络(Fully Convolutional Network, FCN) [4]被提出解决图像分割问题以来，深度学习技术在准确率和效率方面开始超越传统的基于手工特征[5]方法。一些创新模型如 Deeplab [6]、PSPnet [7]有效地学习像素间的上下文关系，显著提高了分割的准确性。为了进一步提升性能，研究人员开发了多种策略，使这些模型能够捕获丰富的上下文信息，同时保留关键细节。然而，这些模型往往伴随着高昂的计算成本，这一点在需要实时处理的应用场景中如自动驾驶[8]和机器人辅助手术[9]中不能满足需求。因此，尽管这些模型取得了卓越的分割性能，但在实际应用中的普及仍面临挑战。

为了满足实时分割的需求，研究者们已经提出了许多高速语义分割的模型。ENet [10]在早期采用了轻量级解码器，对特征图进行了下采样。ICNet [11]对小尺寸输入进行复杂的深度编码，以解析高级语义。MobileNets [12]用深度可分离卷积取代了传统的卷积。这些早期的工作减少了分割模型的延迟和内存使用，但低准确率显著限制了它们在实际应用的普及。近年来，许多基于双分支和多分支的模型被提出，随着移动设备部署需求的不断增长实现了较好的分割速度和精度之间的平衡。

在本文中，我们提出了一个具有深度高分辨率表示的三分支网络用于道路驾驶图像的实时语义分割。我们的网络从一个主干开始，然后分成三个具有不同分辨率的平行分支。第一个分支生成相对高分辨率的特征图提取空间信息，第二个分支通过使用多次残差 U 块提取多尺度语义信息，第三个分支生成边界信息，在三个分支之间架起多个双边连接，实现信息融合。此外，我们引入了上下文变换器模块(Contextual Transformer, CoT)，该模块融合卷积神经网络(CNN)的局部感知能力和 Transformer 的全局依赖建模能力，可以提高模型对图像中不同部分之间关系的理解和表示能力。

本文主要贡献如下：

(1) 使用 U2-Net [26]中提出的 RSU 模块来构建网络的上下文信息分支，大大提高了网络的多尺度特征提取能力。

(2) 在下采样阶段使用 CoT 模块[25]，生成临近间上下文信息和全局上下文信息，从而增强模型的视觉表达能力。

(3) 构建了一个实时的三分支语义分割网络，没有使用预训练的情况下在 Cityscapes 数据集上以 76.5

FPS 的速度取得了 78.6%的 IMoU。

2. 相关工作

实时语义分割网络由于其在实际应用中的需求日益增长而备受关注。为了在推理速度和准确率之间取得平衡，许多研究都做出了贡献。下面按照三个方面来介绍以前的工作。

2.1. 基于编码器–解码器结构的语义分割

一些方法选择了经典的编码器–解码器架构作为其主要框架。这些方法通过逐层下采样和特征融合操作来提取语义特征，有效地结合了低级细节和高级语义信息。例如，SGCPNet [1]提出了空间细节引导的上下文传播策略，在解码器阶段利用空间细节来引导低分辨率全局上下文的传播，有效地重建丢失的空间信息。ESPNet [19]采用了不同扩张率的并行卷积，以此来扩大感受野，提升解码器的效率。EDANet [20]提出的 EDA 模块通过在较大的区块内更紧密地连接输入图像和输出特征，实现了在更广泛的接收范围内共享信息。DFANet [21]通过深度多尺度特征聚合和使用轻量级的深度可分离卷积，有效地细化了高级和低级特征。虽然这些方法在实现了较好的分割精度，但它们在处理高分辨率输入图像时，大多数仍面临着较慢的推理速度的挑战。

2.2. 基于多分支结构的语义分割

与编码器–解码器结构方法不同，多分支结构通过独立提取不同尺度的特征来保留网络早期提取的高分辨率细节以及网络后期提取的上下文信息。BiSeNet [13]提出了一种由上下文分支和空间信息分支组成的双分支架构。上下文分支基于一个预训练主干提取上下文信息，空间信息分支利用几个卷积层来关注空间细节。此后，BiSeNetV2 [17]进一步简化了网络结构。提出双边引导聚合取代 BiSeNet [13]中的特征融合模块，设计了完全手工化的语义分支。所有这些努力都提高了网络的效率。然而，分支之间缺乏有效的特征交互，部分导致了准确率的下降，所有的特征都在网络的末端融合，形成了新的瓶颈。因此，Fast-SCNN [14]和 DDRNet [16]采用了共享骨干网架构。它们从一个分支开始构建网络，然后分成两个具有不同分辨率的平行深分支。这种设计在网络早期共享部分网络参数，允许他们在网络中间添加许多交互。PIDNet [15]建立与 PID 控制器的联系提出了三分支结构，三个分支来分别解析细节、上下文和边界信息，并利用边界关注来指导细节和上下文分支的融合。实现了推理速度和精度之间的最佳权衡。

2.3. 基于注意力机制的语义分割

注意机制用于解决神经网络的局部问题，它可以将局部信息与全局信息联系起来，选择需要更多注意的信息。SENet [28]使用了通道注意力显式建模通道之间的相互依赖性来自适应地重新校准通道方面的特征响应，提高了网络的特征表示能力。DANet [22]使用了双重注意力，使用位置注意模块来聚焦空间信息，使用通道注意模块来关联通道信息。Jin 等人[24]提出注意力引导多模态交叉融合模块(ACFM)，在多阶段利用融合的增强特征表示。此外，受自注意力在各种自然语言处理任务中不断取得成绩的启发，研究界开始更多地关注视觉场景中的自我关注。FANet [23]提出了快速空间注意力，这是对自注意力机制的一种简单而有效的修改，通过改变操作顺序，以很小的计算成本捕获相同的丰富空间上下文。CoTNet [25]使用了上下文变换器，充分利用输入键之间的语义信息来引导动态注意矩阵的学习，促进了远程依赖关系的建模，从而增强了视觉表征能力。

3. 方法

在本节中，将详细说明我们工作的细节。图 1 解释了我们网络的构造。每个组件的详细描述如下。

3.1. 语义分割模型整体架构

图 1 是我们提出的网络模型框架总体结构，首先，通过一个共用的快速下采样阶段将输入图像下采样 8 倍，然后分别引出三个分支，空间信息分支、上下文信息分支、边界信息分支。空间信息分支通过残差基本块和残差瓶颈块提取图像的空间特征，强调物体的形状和大小等细节信息。上下文信息分支使用了残差 U 块来获取多尺度特征，以及利用上下文变换器模块(CoT)来增强全局依赖建模的能力，这对于理解图像中的整体结构和关系至关重要。边界信息分支则侧重于提取精确的边界信息，这对于在复杂场景中区分相邻的物体非常有用。在网络结构中，不同分辨率的特征图通过级联(Concat)操作进行合并，并在合并前后都使用 CoT 模块处理以增强特征表达。此外，网络还引入了多尺度损失计算，通过在不同的分支上附加辅助分类器，并计算与真实标签的损失值，分别对应不同类型的信息和分割任务的需求。这种多尺度监督的方法有助于在训练过程中提供额外的梯度信息，从而促进网络在各个层次上的学习。最后，三个分支在网络的末端进行整合，输出精细的分割结果。通过这种多分支结构设计，旨在有效地结合多种类型的信息，以提高在复杂城市道路场景中的分割任务的性能，确保高精度和高效率的模型运行。接下来，我们将详细描述残差 U 块、上下文变换器模块、损失函数。

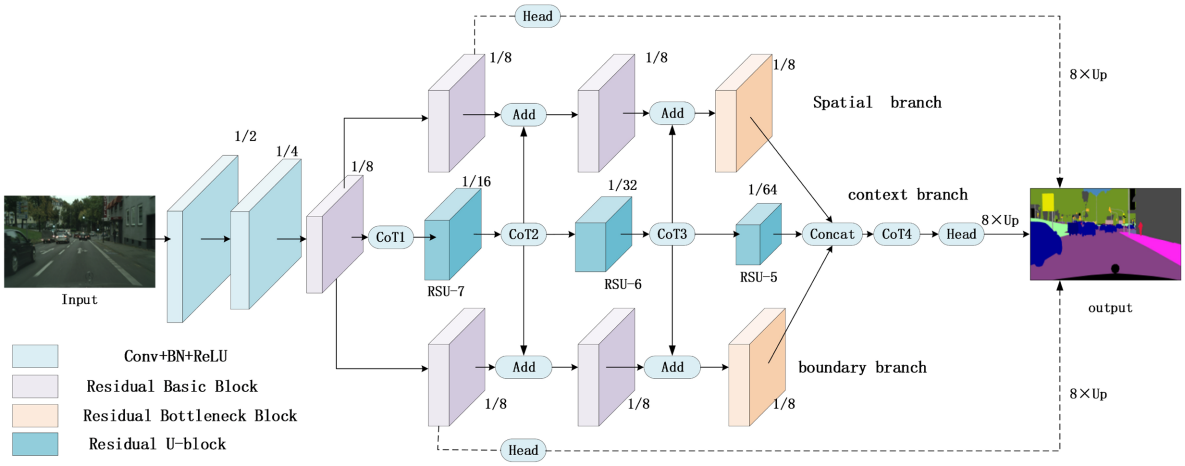

Figure 1. Network architecture overall structure diagram
图 1. 网络架构总体结构图

3.2. RSU 模块

残差 U 块结合了 ResNet 的残差连接思想和 U-Net 的对称扩张路径结构。它由编码器(收缩路径)和解码器(扩张路径)两部分组成，编码器用于捕捉图像的上下文信息，而解码器则用于精确地定位和恢复图像的详细信息。每个残差 U 块内部都包含残差连接，这些连接允许梯度直接流向前面的层，从而缓解了深度网络中梯度消失的问题。

在网络中的上下文信息分支中，我们引入了 U2-Net [26]的剩余 U 块。RSU 模块是高度为 L 的类似 U-Net [27]的对称编码器–解码器结构，RSU-L (L = 7) (C_{in}, M, C_{out}) 的结构如图 2 所示，其中 L 为编码器的层数，C_{in} 为输入通道，C_{out} 输出通道，M 为 RSU 内部各层的通道数。RSU 主要由 3 个部分组成。

(1) 输入卷积层，它将输入特征 $X\left(H\times W\times C_{in}\right)$ 转换为具有 C_{out} 通道的中间映射 $f(x)$，这是一个用于局部特征提取的普通卷积

(2) 高度为 L 的类似 U-Net 的对称编码器-解码器结构，以中间特征映射 $f(x)$为输入，学习提取和编码多尺度上下文信息 $U\left(f\left(x\right)\right)$。$U$ 表示 U-Net 型结构，L 越大，剩余 U 块越深，池化操作越多，接受域

范围越大，局部和全局特征越丰富。

(3) 残差连接，通过求和 $F(x)+U(f(x))$ 融合局部特征和多尺度特征。对于任意 RSU 尺度，输入特征映射和输出特征映射具有相同的分辨率。在上下文信息分支中，我们线性堆叠了 3 个不同大小的 RSU 模块。因此，我们的上下文信息分支由 3 个阶段组成，每个下采样阶段由一个配置良好的 RSU 块填充，L 的尺度为 $(7,6,5)$。每个 RSU 块后连接一个步幅为 2 的最大池化层进行下采样。

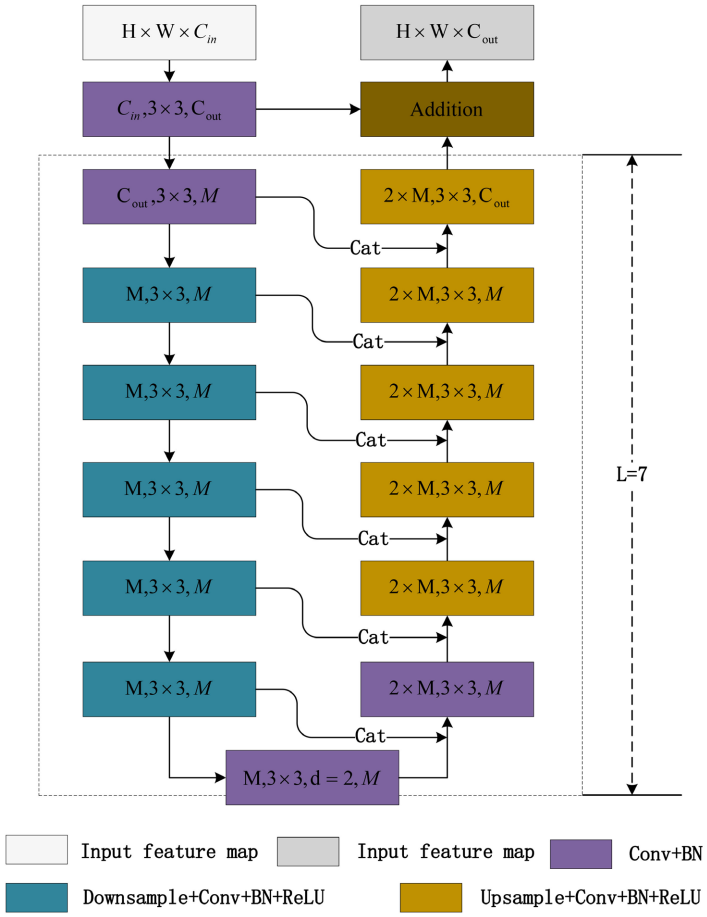

Figure 2. Residual U-block structure diagram
图 2. 残差 U 块结构图

3.3. 上下文变换器模块

CoT 模块是对 Transformer 的改进，传统的 Transformer 结构通过直接在特征图上使用自注意力来获得基于每个空间位置的独立查询键对的注意力矩阵，但未充分利用相邻键之间的丰富的上下文信息，这限制了在二维特征图上的自注意学习能力。并且只使用卷积的话全局建模能力又受到限制。因此引入了 CoT 模块，该模块既结合自注意力机制的优点，又利用了相邻键之间的丰富上下文信息。CoT 模块的结构如图 3 所示，首先通过 $k \times k$ 的卷积操作对输入键的局部信息进行提取，这样就可以得到 Key 邻近键的静态上下文信息 $K1$。具体操作如公式(1)所示：

$$K1 = \mathrm{Re}LU\left(BN\left(Conv_{k\times k}(x)\right)\right) \tag{1}$$

其中 ReLU 表示 ReLU 激活函数，BN 表示批量标准化，$Conv_{k\times k}$ 表示卷积核为 k 的卷积操作。X 为输入

特征。

随后键 Key 经过编码后生成的静态上下文特征信息 $K1$ 与输入查询 Query 连接，并通过两个连续的 1 × 1 卷积获取动态注意力矩阵 $\boldsymbol{A}$，具体操作如公式(2)所示：

$$\boldsymbol{A}=\left[K1,Q\right]W_{\theta}W_{\delta} \tag{2}$$

然后将学到的注意力矩阵 $\boldsymbol{A}$ 乘以输入值 value 以实现输入特征图的动态上下文表示 $K2$。具体操作如公式(3)和(4)所示：

$$V=BN\left(Conv_{1\times1}\left(x\right)\right) \tag{3}$$

$$K2=V*A \tag{4}$$

最后将静态上下文信息 $K1$ 与动态上下文信息 $K2$ 融合作为最终输出。这种设计结构充分利用了 Key 的上下文信息，增强了模型的视觉表达能力。最后通过邻近键上下文信息与全局上下文信息的融合得到输出结果。具体如公式(5)所示：

$$K=K1+K2 \tag{1}$$

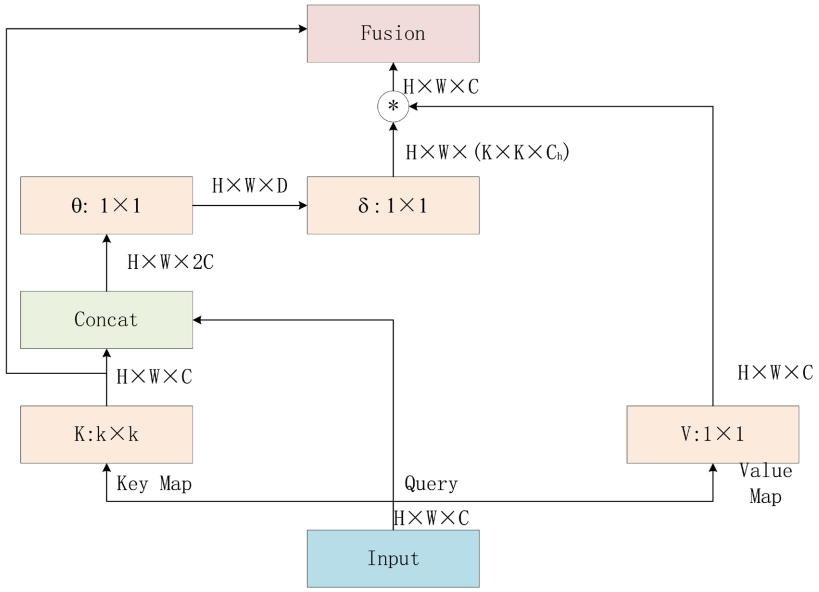

Figure 3. Contextual transformer structure diagram
图 3. 上下文变换器结构

3.4. 损失函数

本模型中采用了一种复合损失函数，空间分支的第一个特征融合后的输出处放置了一个语义分割头，采用交叉熵损失函数 l_0 对空间信息分支的输出进行优化，从而更好地加强整个网络的空间信息特征。对上下文信息分支的输出采用交叉熵损失 l_1 进行优化，来获取更精确的上下文信息。在边界信息分支采用加权二元交叉熵损失 l_2 来处理边界不平衡问题，因为对于小物体，粗糙的边界更容易突出边界区域。最后利用 l_3 的边界感知交叉熵损失和边界头的输出来协调语义分割和边界检测任务。损失函数计算可以写成如(6)所示：

$$l_0=l_1=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}y_{i,c}\log\hat{y}_{i,c} \tag{6}$$

其中 N 是样本数量，C 为总类别数，$y_{i,c}$ 是一个指示变量，表示样本 i 是否属于类别 c (如果像素 i 属于类别 c，则 $y_{i,c}=1$,否则，$y_{i,c}=1$， $\hat{y}_{i,c}$ 表示模型预测 i 像素属于 c 类的概率。

$$l_2=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\right] \tag{7}$$

其中 N 是样本数量，y_i 是第 i 个像素的真实标签， $\hat{y}_i$ 表示模型预测 i 像素为正类的概率。

$$l_3=\sum_{i,c}\{b_i>t\}\left(\mathrm{y}_{\mathrm{i,c}}\log y_{i,c}\right) \tag{8}$$

其中 t 为一个预定义阈值，本文中设置成 0.8，b_i 表示像素 i 为 c 类的边界头输出，$y_{i,c}$ 表示第 i 个像素为 c 类的分割标签， $\hat{y}_{i,c}$ 为上下文信息分支分割第 i 个像素为 c 类的预测结果。

$$\mathrm{Loss}=\mu_0 l_0+\mu_1 l_1+\mu_2 l_2+\mu_3 l_3 \tag{9}$$

公式(11)表示模型总体的损失函数，其中 $\mu_0=0.4,\mu_1=\mu_2=1,\mu_3=20$。

4. 实验

我们在Cityscapes和CamVid数据集上进行了实验。首先介绍数据集和实现细节。接下来，在Cityscapes验证集以及 CamVid 测试集上，将本文提出的方法与其他方法进行了比较。

4.1. 数据集和验证指标

Cityscapes 是侧重于从驾驶角度对城市街道场景的语义理解。将该数据集的精细标注图像分为训练集、验证集和测试集，分别为 2975 张、500 张和 525 张。所有的图像都是 1024 × 2048 的分辨率。在我们的实验中，我们按照 19 个语义类别标准，而不是原来的 34 个类别。我们只使用精细标注的图像来训练和验证我们的网络。

CamVid (Cambridge-drivingLabeled Video Database)也是一个用于实时语义分割的街景数据集，它包含 701 个密集标注的图像，每个图像的分辨率为 720 × 960。这些图像分为 367 个训练图像，101 个验证图像和 233 个测试图像。CamVid 有 32 个类别，其中有 11 个类别的子集用于分割任务。

对于实验中涉及的所有方法，以及本文提出的方法，我们都采用了均交并比(MIoU)和每秒帧数(FPS)作为评价指标。

4.2. 实验设置

由于我们的模型没有使用预训练，我们需要在 Cityscapes 训练集上从头开始训练我们的模型。我们选择的优化器是 SGD 算法，动量设置为 0.9。在 Cityscapes 数据集上，我们随机裁剪输入图像到 0.5 到 2 倍分辨率大小，批处理大小设置为 12，对于学习率，我们使用 0.01 作为初始设置，$5e^{-4}$ 作为优化器中的权重衰减。训练轮数的设置为 600，以使模型能够充分学习。在 CamVid 数据集上，批量大小设置为 12。由于缺乏足够的训练图像，我们设置 0.005 为初始学习率，$5e^{-4}$ 为权值衰减。此外，使用在 Cityscapes 预训练的模型，训练轮数设置为 300，以避免训练集上的过拟合。

4.3. 对比实验

在上面描述的实验配置中，我们将我们的网络与以前的实时方法进行比较。所有在 Cityscapes 上评估的方法都使用 1024 × 2048 的分辨率。表 1 显示了每种方法的模型信息，包括 mIoU、GPU 设备和 FPS。其中一些方法我们遵循其原始训练设置并评估其在我们设备上的性能，用符号*标记它们。其他数据均摘自上述方法的原始论文。实验结果表明，我们的网络具有较好的效果。

Table 1. Cityscapes data set results
表 1. Cityscapes 数据集结果

模型	GPU	MIoU	PFS
BiSeNet V2	RTX 1080Ti	73.4	156
BiSeNet V2-L	RTX 1080Ti	75.8	47.3
Fast-SCNN*	V100	68.3	125
DDRNet-23-slim*	V100	76.6	104.3
PIDNet-S*	V100	77.8	103.5
本文方法	V100	78.6	76.5

在 CamVid 数据集上对我们的方法进行了测试(见表 2)。所有在 CamVid 上评估的方法都使用了 960 × 720 的原始分辨率，与表 1 相同，符号*表示我们在设备和环境设置上复现的结果。在 CamVid 数据集上的实验结果表明，我们的方法也取得了很好的性能。

Table 2. CamVid data set results
表 2. CamVid 数据集结果

模型	GPU	mIoU	PFS
BiSeNet V2	RTX 1080Ti	76.7	124.5-
BiseNet V2-L	RTX 1080Ti	78.5	32.7
DDRNet-23-slim*	V100	76.3	226.6
PIDNet-S*	V100	76.8	199.4
本文方法	V100	77.6	145.9

4.4. 可视化结果

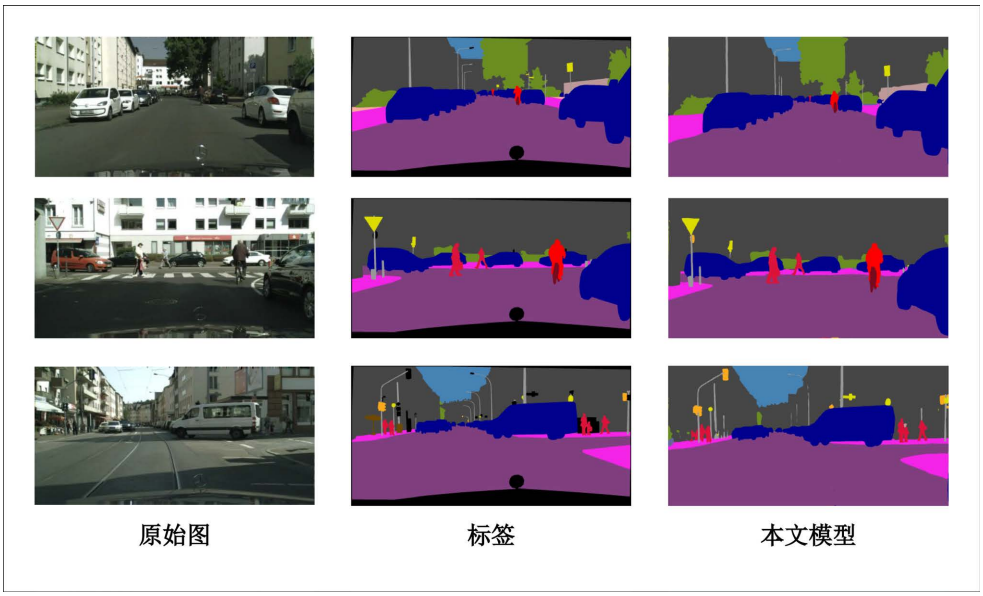

Figure 4. Visual result
图 4. 可视化结果

本文提出的模型在 Cityscapes 上的分割效果如图 4 所示。三列图像是分别是初始图，标签图和模型的分割效果图。在本文模型输出的分割图上对不参与分割的类进行了屏蔽。

5. 结束语

本文提出了一种基于残差 U 块和上下文变换器的三分支实时语义分割网络，在保证分割准确度的同时兼顾了实时性，上下文路径的 RSU 模块和 CoT 模块可以获取丰富的上下文信息以及建立全局依赖。实验表明我们的方法在两个流行的基准测试中取得了具有竞争力的结果。

参考文献

[1] Feng, D., Haase-Schütz, C., Rosenbaum, L., *et al*. (2020) Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Transactions on Intelligent Transportation Systems*, **22**, 1341-1360. https://doi.org/10.1109/TITS.2020.2972974

[2] Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., *et al*. (2021) Deep Semantic Segmentation of Natural and Medical Images: A Review. *Artificial Intelligence Review*, **54**, 137-178. https://doi.org/10.1007/s10462-020-09854-1

[3] Yuan, X., Shi, J. and Gu, L. (2021) A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery. *Expert Systems with Applications*, **169**, Article ID: 114417. https://doi.org/10.1016/j.eswa.2020.114417

[4] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *The Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[5] Shotton, J., Johnson, M. and Cipolla, R. (2008) Semantic Text on Forests for Image Categorization and Segmentation. *Proceedings of the* 2008 *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 23-28 June 2008, 1-8. https://doi.org/10.1109/CVPR.2008.4587503

[6] Chen, L.-C., Papandreou, G., Kokkinos, I., *et al*. (2014) Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected Crfs.

[7] Zhao, H., Shi, J., Qi, X., *et al*. (2017) Pyramid Scene Parsing Network. *The Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 6230-6239. https://doi.org/10.1109/CVPR.2017.660

[8] Xiao, X., Zhao, Y., Zhang, F., *et al*. (2023) BASeg: Boundary Aware Semantic Segmentation for Autonomous Driving. *Neural Networks*, **157**, 460-470.

[9] Shvets, A.A., Rakhlin, A., Kalinin, A.A., *et al*. (2018) Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning. *Proceedings of the* 2018 17*th IEEE International Conference on Machine Learning and Applications* (*ICMLA*), Orlando, 17-20 December 2018, 624-628. https://doi.org/10.1109/ICMLA.2018.00100

[10] Paszke, A., Chaurasia, A., Kim, S., *et al*. (2016) Enet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation.

[11] Zhao, H., Qi, X., Shen, X., *et al*. (2018) Icnet for Real-Time Semantic Segmentation on High-Resolution Images. *The Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 418-434. https://doi.org/10.1007/978-3-030-01219-9_25

[12] Howard, A.G., Zhu, M., Chen, B., *et al*. (2017) Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[13] Yu, C., Wang, J., Peng, C., *et al*. (2018) Bisenet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. *The Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 334-349. https://doi.org/10.1007/978-3-030-01261-8_20

[14] Poudel, R.P., Liwicki, S. and Cipolla, R. (2019) Fast-Scnn: Fast Semantic Segmentation Network.

[15] Xu, J., Xiong, Z. and Bhattacharyya, S.P. (2023) PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. *The Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 17-24 June 2023, 19529-19539. https://doi.org/10.1109/CVPR52729.2023.01871

[16] Hong, Y., Pan, H., Sun, W., *et al*. (2021) Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Road Scenes.

[17] Yu, C., Gao, C., Wang, J., *et al*. (2021) Bisenet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *International Journal of Computer Vision*, **129**, 3051-3068. https://doi.org/10.1007/s11263-021-01515-2

[18] Hao, S., Zhou, Y., Guo, Y., *et al*. (2022) Real-Time Semantic Segmentation via Spatial-Detail Guided Context Propagation. *IEEE Transactions on Neural Networks and Learning Systems*.

[19] Mehta, S., Rastegari, M., Caspi, A., *et al*. (2018) Espnet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. *The Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 561-580. https://doi.org/10.1007/978-3-030-01249-6_34

[20] Lo, S.-Y., Hang, H.-M., Chan, S.-W., *et al*. (2019) Efficient Dense Modules of Asymmetric Convolution for Real-Time Semantic Segmentation. *Proceedings of the* 1*st ACM International Conference on Multimedia in Asia*, Beijing, 15-18 December 2019, 1-6. https://doi.org/10.1145/3338533.3366558

[21] Li, H., Xiong, P., Fan, H., *et al*. (2019) Dfanet: Deep Feature Aggregation for Real-Time Semantic Segmentation. *The Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 9514-9523. https://doi.org/10.1109/CVPR.2019.00975

[22] Fu, J., Liu, J., Tian, H., *et al*. (2019) Dual Attention Network for Scene Segmentation. *The Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 3141-3149. https://doi.org/10.1109/CVPR.2019.00326

[23] Hu, P., Perazzi, F., Heilbron, F.C., *et al*. (2020) Real-Time Semantic Segmentation with Fast Attention. *IEEE Robotics and Automation Letters*, **6**, 263-270. https://doi.org/10.1109/LRA.2020.3039744

[24] 靳瑜昕, 杨晓文, 张元, 等. 注意力引导多模态融合的 RGB-D 图像分割[J]. 计算机工程与设计, 2022, 43(12): 3453-3460.

[25] Li, Y., Yao, T., Pan, Y., *et al*. (2022) Contextual Transformer Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 1489-1500. https://doi.org/10.1109/TPAMI.2022.3164083

[26] Qin, X., Zhang, Z., Huang, C., *et al*. (2020) U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. *Pattern Recognition*, **106**, Article ID: 107404. https://doi.org/10.1016/j.patcog.2020.107404

[27] Li, X., Chen, H., Qi, X., *et al*. (2018) H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. *IEEE Transactions on Medical Imaging*, **37**, 2663-2674. https://doi.org/10.1109/TMI.2018.2845918

[28] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *The Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745