文章编号：2095-2163(2023)08-0011-07 中图分类号：TP391 文献标志码：A

基于 FuseNet 的多模态融合图像分割网络

张 涛，黄孝慈
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘 要：现有的图像分割工作主要是利用 CNN 学习高层语义特征中的上下文信息，直接生成最终的分割模板，没有对图像中类别信息进行显式建模。因此，本文提出了基于 FuseNet 的多模态融合图像分割网络，其目的是在输入图像中通过语言表达生成对象的分割图像。模型由 3 个核心部件组成，分别是：多模态融合模块、定位模块和 Segmentation Mask 模块。视觉特征与语言特征的融合可以关注输入语言中多个指定类别的目标图像区域，然后根据对象的上下文生成一个关于对象的精细分割图像。实验结果表明，本文方法在真实环境中获得了最佳的 *mIoU* 值（52.1%），比仅在源数据上训练的模型增加了 15.5%。通过模型在数据集上的性能评估，利用对视觉和语言特征的显式建模，由此得到了比先前模型更精确的分割结果和更快的分割性能。

关键词：图像分割；深度学习；多模态融合；FuseNet

Multimodal fusion image segmentation network based on FuseNet

ZHANG Tao, HUANG Xiaoci
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 The purpose of localization image segmentation is to segment objects corresponding to natural language expressions. Previous methods usually focus on designing an implicit and recursive feature interaction mechanism to fuse visual-linguistic features and directly generate the final segmentation template without explicitly modeling the localization information of reference instances. Therefore, this paper proposes a multimodal fusion image segmentation network based on FuseNet, which aims to generate segmented images of objects through linguistic expressions in the input images. The model consists of three core components, which are a multimodal fusion module, a localization module and a Segmentation Mask module. The fusion of visual features and linguistic features can focus on multiple target image regions of specified categories in the input language and then generate a finely segmented image about the object according to its context. The performance of the proposed method is evaluated on the Cityscapes dataset with real-world image segmentation and containing 50 different urban street scenes. The experimental results show that the best *mIoU* value (52.1%) is increased by 15.5% in the real environment models trained on the data. Through the performance evaluation of the model on the dataset, more accurate segmentation results and faster segmentation performance are obtained than previous models using explicit modeling of visual and linguistic features.

【Key words】 image segmentation; deep learning; multimodal fusion; FuseNet

0 引 言

近年来，基于深度学习的图像分割方法[1-3]因其具备的精心设计框架，以及各种细分数据集的可用性已取得了很大进展。其中，来自各种深层网络学习到的更好的特征表示对该方法的迅猛发展发挥了至关重要的核心作用。然而，对于许多现实世界的应用、例如医疗和制造业，收集和标记数据非常耗时，需要用到专业的注释员。这个问题的直观解决方法是在现有模型的源数据集上训练未标记目标域。然而，由于源域和目标域中的各种数据分布而导致的域转移问题往往会阻碍该解决方法的实现。此外，方法在实现过程中没有在语言表达的指导下明确定位参考对象，只利用耗时的后处理 DCRF 生成最终的细化分割。对于开放集[4-5]图像分割任务，现已获得了广泛的应用，例如交互式图像编辑和语言引导的人机交互。除了传统的图像分割，由于图像和语言之间的语义差异，语言相关的图像分割更具挑战性。此外，文本表达不仅限于实体（例如，“人”、“马”），还可能包含描述性词语，如对象属性

作者简介：张 涛（1995-），男，硕士研究生，主要研究方向：深度学习、智能网联汽车。
通讯作者：张 涛 Email：2099743507@qq.com

收稿日期：2022-09-05

(例如“红色”、“年轻”)、动作(例如“站立”、“保持”)。

以前的研究主要集中在如何融合图像特征和语言特征。一个简单的解决方案[6]是利用串联和卷积的方法融合视觉和语言表达,以产生最终的分割结果。但是,由于视觉和文本信息是单独建模的,这种方法不能有效地建模图像和语言之间的对齐。为了进一步模拟多模态特征之间的上下文,一些先前的方法[7]提出了跨模态注意,自适应地关注图像中的重要区域和语言表达中的信息关键词。最近,Hu 等学者[8]利用卷积神经网络(convolutional neural networks ,CNNs)和长-短期记忆网络(long short-term memory, LSTM)[9]的视觉和语言特征串联来生成分割模板。为了获得更精确的结果,文献[10]融合了多层次的视觉特征,以细化分割掩模的局部细节。

综上所述,尽管这些方法都已获得了长足的发展,但网络体系结构和实验实践却已逐步变得更加复杂。这也导致算法的分类与比较显得更加困难。因此,针对这一现状,研究中从另一个角度考虑解决这个问题。这里将图像分割任务分解为 2 个子序列任务,分别是:词向量特征提取和精细分割掩模生成。在本文提出的模型中,主要由以下核心部件组成:

(1)多模态融合模块。视觉特征和语言特征分别由卷积神经网络(SegNet)和 LSTM 网络提取,然后融合生成多模态特征。

(2)定位模块。使用基于注意力机制构建的 transformer 将会自适应地获取图像中的重要区域和语言表达中的信息关键词之间的相关性。

(3)Segmentation Mask 模块。使用多采样率和有效卷积特征层,从而在多尺度上捕获对象和图像上下文,并将反卷积特征图的采样率提高,由此获得更精确的分割结果。最后,使用交叉熵损失函数训练网络。

1 FuseNet 算法基础

1.1 语言特征提取

给定一个背景词向量 $X=[x_1,x_2,\cdots,x_m]$,其中 x_i 是第 i 个标记。首先应用表查找来获得单词嵌入,之后将其初始化为一个 300 维的通道嵌入向量,每个通道表示一个词向量的维度,再通过 GLOVE 进行输入[11]。为了模拟相邻单词之间的相互依赖关系,使用标准的 LSTM 来处理初始嵌入文本向量:

$$\boldsymbol{h}_{t1}=LSTM(\boldsymbol{x}_t,\boldsymbol{h}_{t-1}),\ \boldsymbol{h}_0=0 \tag{1}$$

$$\boldsymbol{h}_{t2}=LSTM(\boldsymbol{x}_t,\boldsymbol{h}_{t+1}),\ \boldsymbol{h}_{m+1}=0 \tag{2}$$

其中,$\boldsymbol{h}_{t1}$ 和 $\boldsymbol{h}_{t2}$ 分别表示 LSTM 向前和向后获得的文本向量。全局文本通过所有单词之间的平均池化获得,其定义如下:

$$\boldsymbol{p}_{text}=avg(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_m) \tag{3}$$

$$\boldsymbol{h}_t=concat(\boldsymbol{h}_{t1},\boldsymbol{h}_{t2})\quad t\in[1,2,\cdots,m] \tag{4}$$

1.2 视觉特征提取

给定输入图像 $I\in\mathbb{R}^{H\times W\times 3}$,利用视觉主干提取多级视觉特征,即 $F_{e_1}\in\mathbb{R}^{\frac{H}{64}\times\frac{W}{64}\times d_1}$,$F_{e_2}\in\mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times d_2}$ 和 $F_{e_3}\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times d_3}$。这里,$H$ 是原始图像的高度,W 是原始图像的宽度,d 是特征通道的尺寸。对于图像中的每个像素,$\{1,\cdots,P_{point}\}^{\frac{H}{64}\times\frac{W}{64}\times d_1}$。研究假设这些像素对应于场景中的静态部分,即图像中的背景变化仅由相机运动引起。将最终卷积层所获得的视觉特征通过 MLP 反向投影成高维 3D 像素点,有利于像素分类并用于后续的定位环节。3D 像素点投影如图 1 所示。

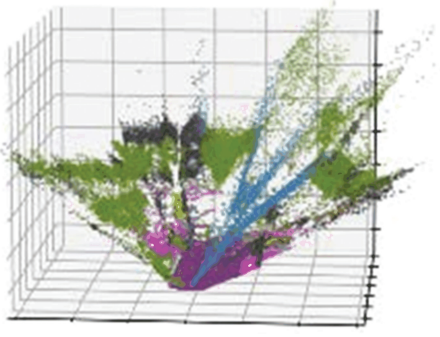


图 1 3D 像素点投影

Fig. 1 3D pixel projection

2 FuseNet 总体架构

整体模型架构如图 2 所示,本文中模型的输入由图像 I 和背景词向量 X 组成。为了模型的轻量化,解码器模块具有相对于编码器模块的对称结构,其中输入和输出通道的数量相反。研究中,使用 SegNet 和 LSTM 分别提取 I 和 X 的特征,随后送入多模态融合模块,融合生成多模态特征。其次,使用基于注意力机制构建的 transformer 将会自适应地获取图像中的重要区域和语言表达中的信息关键词之间的相关性。最后,使用多采样率和有效卷积特征层,有利于在多尺度上捕获对象和图像上下文,并使反卷积特征图的采样率得以提升,从而获得更精确的分割结果。

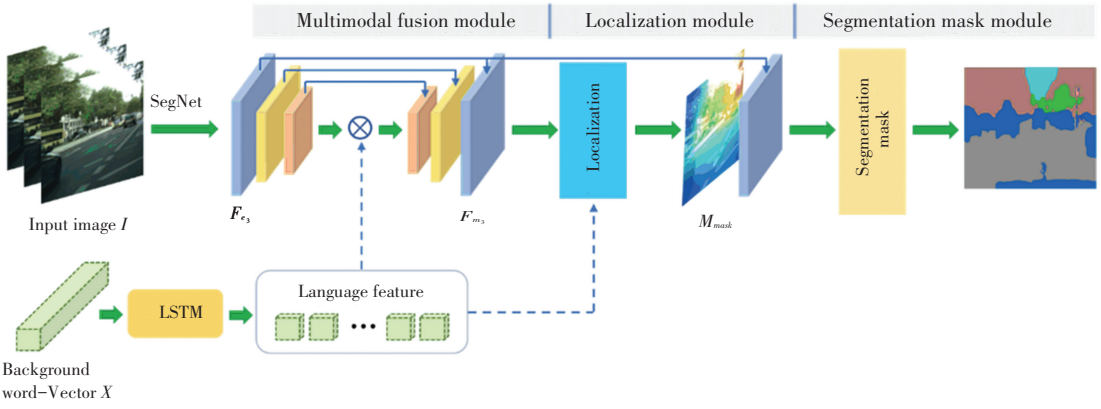


图 2　整体模型架构

Fig. 2　Overall model architecture

2.1　多模态融合模块

由图 2 可知，研究中通过融合 F_{e_1} 和 P_{text} 获得多模态张量，公式如下：

$$\boldsymbol{f}^{l}_{m1} = g(\boldsymbol{f}^{l}_{e1}\boldsymbol{W}_{e\ 1}) \cdot g(\boldsymbol{p}_{text}\boldsymbol{W}_{t}) \tag{5}$$

其中，g 表示 *ReLU* 激活函数；$\boldsymbol{f}^{l}_{m_1}$ 和 $\boldsymbol{f}^{l}_{e_1}$ 分别是 $\boldsymbol{F}_{m_1}$ 和 $\boldsymbol{F}_{e_1}$ 的特征向量；$\boldsymbol{W}_{e_1}$ 和 $\boldsymbol{W}_t$ 是将视觉和词文本表示转换为相同特征维度的 2 个转换矩阵。然后，多模态张量 $\boldsymbol{F}_{m_2}$ 和 $\boldsymbol{F}_{m_3}$ 通过以下方式获得：

$$\boldsymbol{F}^{'}_{m\ \mu-1} = UpSample(\boldsymbol{F}_{m_{\mu-1}}) \tag{6}$$

$$\boldsymbol{F}_{m_\mu} = concat(g(\boldsymbol{F}^{'}_{m\ \mu-1}\boldsymbol{W}_{m_{\mu-1}}),\ g(\boldsymbol{F}_{e_\mu}\boldsymbol{W}_{e_\mu})) \tag{7}$$

其中，$\mu \in [2,3]$，上采样的步长为 2×2。在下面的过程中，使用 $\boldsymbol{F}_{m_3}$ 作为输入来生成分割掩码。以往的研究通常采用多次注意力机制来获得分割结果。在本文中，先是根据词向量进行定位、再做分割，可以取得良好的性能，对此将展开研究论述如下。

2.2　定位模块

在多模态任务中，一个主要的挑战是建立图像和文本之间的关系模型。近年来，注意力机制已成为功能强大的一种优秀技术，可以在图像分割中提取与语言表达相对应的视觉内容。特征 $\boldsymbol{F}_{m_3}$ 包含丰富的多模态信息，必须进一步建模以获得图像中的相关区域。定位的目的是为了将每个像素与语言表达所涉及的全局分布的视觉区域关联起来，这些区域的反应分数高于不相关区域，用于增强全方位推理，同时防止模型过度拟合图像。研究中将全局文本 $\boldsymbol{P}_{text}$ 视为编码器输出，解码器遵循变压器的标准架构，使用多头注意力机制将多模态特征 $\boldsymbol{F}_{m_3}$ 转换为一个粗略的分段掩码热图 M_{mask}，因此可得：

$$M_{mask} = decoder(\boldsymbol{F}_{m3}, \boldsymbol{P}_{text}) \tag{8}$$

其中，响应分数越高的区域就越有可能对应于语言表达（见图 1）。

解码器需要一个序列作为输入，因此可将 $\boldsymbol{F}_{m_3}$ 的空间维度压缩为一维，从而生成 $d \times \frac{HW}{128}$ 特征映射。由于 transformer 架构是置换不变的，就可使用固定位置编码对其进行补充，这些编码被添加到每个注意层的输入中。

2.3　Segmentation Mask 模块

给定由式（8）中生成的视觉对象，Segmentation Mask 模块的目标是生成最终的精细分割掩模。研究中，先将原始多模态特征 $\boldsymbol{F}_{m_3}$ 和视觉对象 M_{mask} 连接起来，并利用分割模块来细化粗分割结果：

其公式定义如下：

$$H_{mask} = Seg(concat(\boldsymbol{F}_{m3}, M_{mask})) \tag{9}$$

其中，Segmentation Mask 模块的主要结构以及分割过程如图 3 所示。Segmentation Mask 模块的卷积特征层使用了多采样率和全局池化的方式，以便于从多尺度上捕获对象特征和图像上下文。请注意，为了获得更精确的分割结果，通过反卷积的方式将特征图的采样率增加了 4 个因子。这样，预测的掩码 $H_{mask} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}}$。

2.4　模型训练

在模型训练期间采用交叉熵损失函数，其定义如下：

$$S_{seg} = \sum_{e=1}^{\frac{H}{4}\times\frac{W}{4}} [(1-g_e)\log(1-p_e) + g_e\log(p_e)] \tag{10}$$

其中，g_e 和 p_e 分别表示下采样中的地面真相掩码和预测掩码 H_{mask} 的元素。

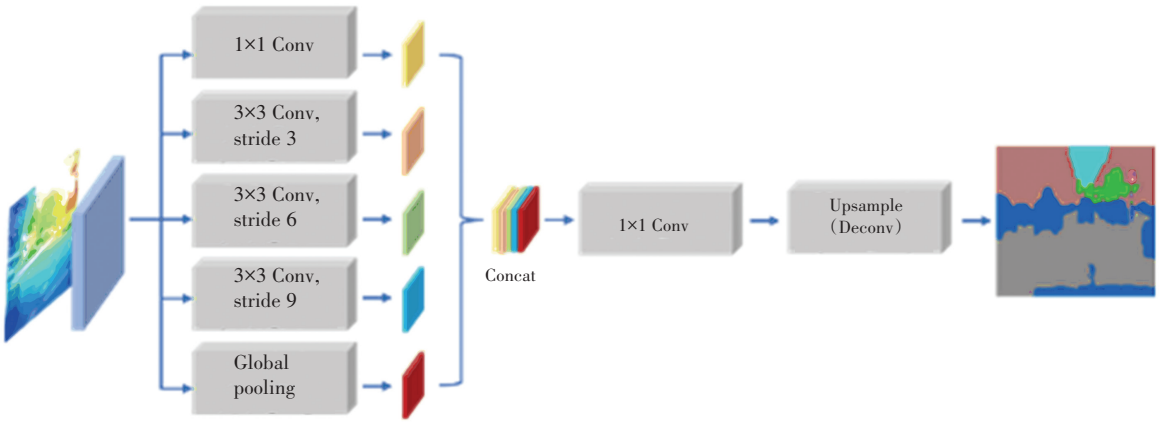


图 3 Segmentation Mask 模块

Fig. 3 Segmentation Mask module

3 实验和结果分析

3.1 数据集

在本小节中,简要介绍用于验证本模型的数据集,即广泛使用的 Cityscapes 数据集[12]。Cityscapes 由 5 000 幅真实的城市交通场景图像组成,分辨率为2 048×1 024,并带有密集像素注释。该数据集中2 975个图像用于培训,500 个图像用于验证,1 525 个图像用于测试。城市景观标注了 33 个类别,其中 19 个用于培训和评估。不含地面真相的训练集用于训练模型,验证集用于评估模型。GTA5[13]是一种合成数据集,其图像从游戏视频中收集,并通过计算机图形技术自动生成相应的语义标签。其中,包括由9 633个像素级标签合成的图像。在 2 种不同的环境下评估了本文提出的 FuseNet 图像分割框架,并按照以前的方法[14],将 Cityscapes 视为目标域,GTA5 视为源域(GTA5-Cityscapes)。

3.2 实施细节

本文使用 Pytork 库实现了提出的方法,并在 NVIDIA 2080TI GPU 上进行了训练。所有网络都使用了随机梯度下降(stochastic gradient descent, SGD)优化器进行训练。初始学习速率和动量分别设置为2.5e-4和 0.9,并采用幂为 0.9 的多项式衰减策略来调整学习速率,接下来将最大迭代次数设置为150 000次。输入图像的大小调整为 416×416,输入句子的最大长度设置为 15。使用 1 024 维的 LSTM 来提取文本特征。过滤维度设置为 1 024。该解码器具有 1 层网络、4 个头和 1 024 个隐藏单元。用平均交集(*mIoU*)来评估本文提出方法的性能。

3.3 定量结果

首先,在 GTA5-Cityscapes 中验证本文方法的有效性,相应的比较结果见表 1。表 1 中,每类的最佳结果以粗体突出显示。从表 1 中可以看出,本文得到的 *mIoU*(52.1%)获得了最佳值,这大大优于其余方法,同时比仅在源数据上训练的模型增加了 15.5%,表现出了优越性能。本文提出的方法在建筑物、墙壁、道路等类别上取得了更显著的改进。这些物体具有刚体,并且在不同的源域中形状相似。*mIoU* 的值越高,也就证明了本文所提出的 Segmentation Mask 模块在学习视觉和语言模态之间语义对齐方面的有效性更强。总地来说,本文提出的分割框架优于其他大部分模型。

表 1 FuseNet 在 GTA5-Cityscapes 上与其他先进模型的对比结果

Tab. 1 Comparison results of FuseNet with other advanced models on GTA5-Cityscapes

Method	road	sidewalk	building	wall	Sign	sky	person	rider	vege	car	fence	*mIoU*
Source only	75.8	16.8	77.2	12.5	30.1	70.3	53.8	26.4	27.9	49.9	41.1	36.6
CRST [14]	91.0	55.4	80.0	33.7	32.9	80.8	57.7	24.6	30.1	84.1	42.3	47.1
MLSL [15]	89.0	45.2	78.2	22.9	**46.1**	61.2	60.4	26.7	44.9	85.4	**49.3**	49.0
UIA [16]	90.6	36.1	82.6	29.5	314.0	80.2	59.3	29.4	**53.9**	86.4	37.6	46.3
Ours	**91.8**	**56.1**	**84.6**	**39.4**	42.9	**84.6**	**65.3**	**37.8**	48.1	**87.5**	48.8	**52.1**

本文收集含有不同类别的图像进行运行时间分析，对比结果如图 4 所示。每次分析重复 400 次，然后取平均值。研究比较了 4 种最先进的方法，包括 Source only、CRST、MLSL、UIA 模型。模型运行时间分析结果如图 4 所示。由图 4 可知，Source only 和 CRST 的推理时间大致与图像中的类数成正比，本文的方法和 MLSL 模型的推理时间与图像中的类数是不变的，并且本文提出的模型比现有的方法快得多。值得注意的是，本文的方法没有使用任何对抗性学习或任何其他复杂的技巧，这可归因于源域组合训练可以在一定程度上提高目标域的性能，源域之间的协作学习比目标域上的协作学习带来了更多的改进。

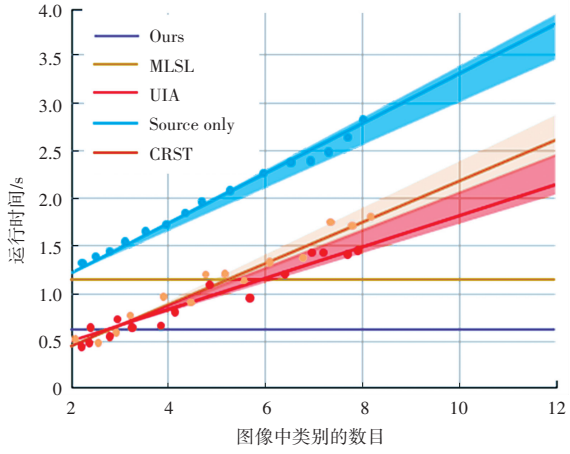


图 4　模型运行时间分析

Fig. 4　Analysis of model running time

图 5 显示了训练过程中分割精度和损失值的变化。2 幅图中的结果可以反映模型随着迭代次数的增加而收敛。如果损失值在几个时期后略有增加，则该模型将被视为收敛条件。在训练过程中经过1 500 次迭代后，该框架达到了收敛条件，并在对比实验中获得了最佳结果，这也验证了表 1 的结论。在第 5 阶段，5 种方法（包括 FuseNet、MLSL、CRST、UIA 和 Source only）的准确度分别为 83.3%、78.2%、65.5%、62.9%和 61.4%。经过1 500个阶段后，本文方法取得了最好的性能并稳定增长，其损失值为−4.61，达到了收敛条件。损失值的变化和最终结果表明，本方法在收敛速度和准确度上优于其他基线方法。

3.4　定性结果

为了直观地评估定性结果，本文提出的基于现有的 MLSL 模型，对含有多类别的图像进行了图像分割，分割结果如图 6 所示。图 6(a)～(c)中，从左至右分别是：Language：马路，车辆，天空，树，标志，墙壁；Language：马路，车辆，行人，树，栅栏，墙壁；Language：马路，车辆，树，天空，墙壁。所有这些图像均来自 GTA5−Cityscapes。从这些定性结果中，可以看到本文的模型根据输入语言所指定的类别对各类型图像都能够以精确分割，所分割出来的事物类型往往是最贴近真值的。本文的模型可以利用依赖于语言和 transformer 中复杂的特征注意力模型，自适应地提取语言表现中的信息关键词，与图片中的重要区域之间的信息关联，从而得到了最匹配的特征分布，加快了推理定位对象的多模态信息融合过程，再通过更精细化的特征分割模块，最后使模型达到了更高的准确度和更好的结构化分割输出。

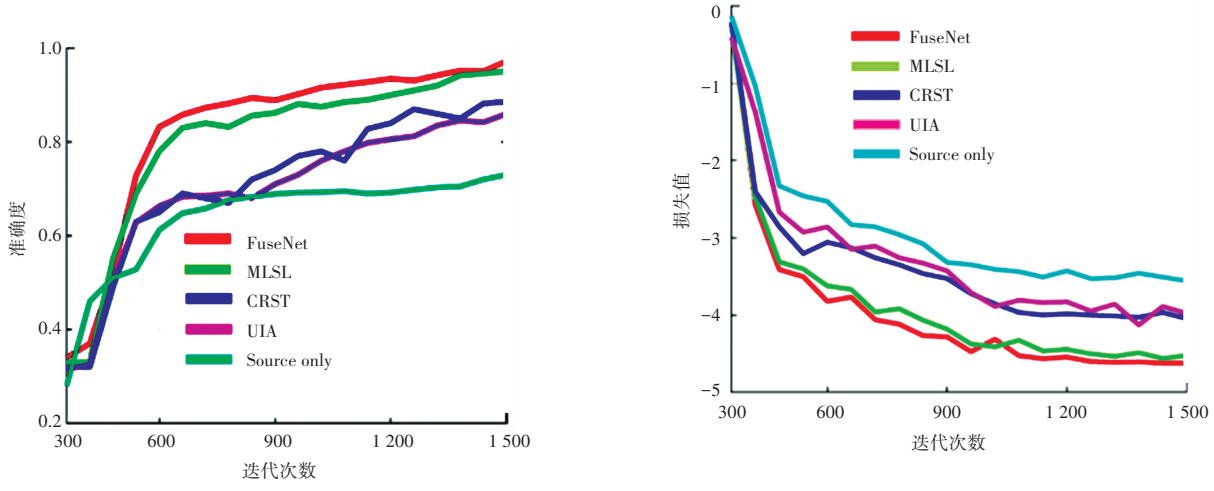
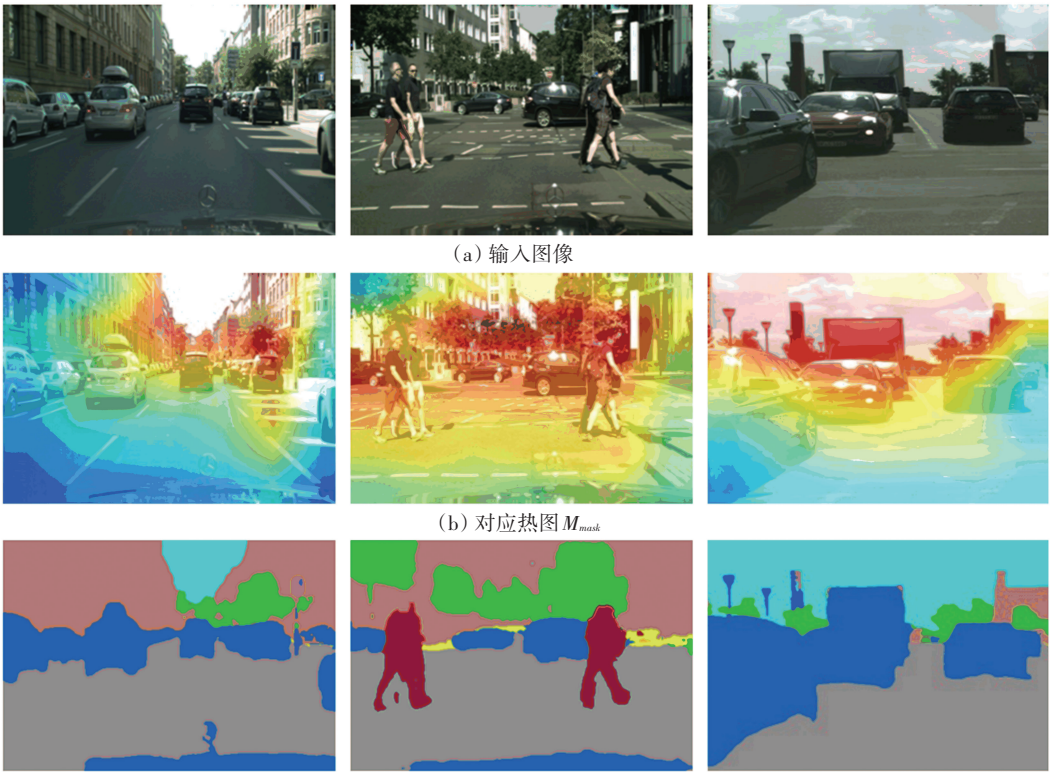


图 5　训练过程中分割精度和损失值的变化

Fig. 5　Change of segmentation accuracy and loss value during training



(a) 输入图像

(b) 对应热图 M_{mask}

(c) 本文的分割结果

图 6 GTA5-Cityscapes 上不同数量的标记目标图像上的定性结果

Fig. 6 Qualitative results of different number of marker target images on GTA5-Cityscapes

4 结束语

在本文中,提出了一种新颖的用于图像分割的自适应框架(FuseNet)。其目的是在输入图像中将语言表达的类别对应的图像进行分割。在研究工作中,为这项任务开发了一种简单而有效的方法。将该任务分解为 2 个子序列任务:词向量特征提取和精细分割掩模生成。首先将提取到的语言和视觉特征送入多模态融合模块,融合生成多模态特征。其次,使用基于注意力机制构建的 transformer 将会自适应地获取图像中的重要区域和语言表达中的信息关键词之间的相关性,用于捕获和传输像素级的语义信息。最后,使用多采样率和有效卷积特征层,从而在多尺度上捕获对象和图像上下文,并将反卷积特征图的采样率提高以获得更精确的分割结果。通过对类别先验的显式建模,减少冗余类别的重复匹配,研究得到了比之前最好的结果更高的分割性能。从上述实验中也证实了本文方法的每个组成部分的有效性。此外,只使用了简单的视觉和语言特征提取主干。更复杂的网络结构有可能进一步提高性能,这将在未来的工作中加以解决。

参考文献

[1] HE Kaiming, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [C]// Proceedings of the IEEE International Conference on Computer Vision. Italy:IEEE, 2017: 2961-2969.

[2] 刘伟伟, 刘金清. 结合 DRLSE 模型的自适应医学图像分割算法[J]. 电子测量技术, 2011, 34(11): 62-65.

[3] 冯林, 孙焘, 吴振宇,等. 基于分水岭变换和图论的图像分割方法[J]. 仪器仪表学报, 2008, 29(03): 649-653.

[4] XIE Enze, SUN Peize, SONG Xiaoge, et al. Polarmask: Single shot instance segmentation with polar representation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 12193-12202.

[5] WANG Xinlong, KONG Tao, SHEN Chunhua, et al. Solo: Segmenting objects by locations[J]. arXiv preprint arXiv: 1912.04488,2019.

[6] HU Ronghang, ROHRBACH M, DARRELL T. Segmentation from natural language expressions [C]// European Conference on Computer Vision. Amsterdam: Springer, 2016, 88 (1/4): 172-145.

[7] CHEN Dingjie, JIA Songhao, LO Y C, et al. See-through-text grouping for referring image [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019, 47(5): 19717-1925.

[8] HU Ronghang, ROHRBACH M, ANDREAS J, et al. Modeling relationships in referential expressions with compositional modular networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA:IEEE, 2017, 25(8): 2220-2233.

(下转第 24 页)