零示例语义分割

计算机科学与技术
周思远

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

目录

目录

研究背景

➢ 语义分割

- 语义分割[1-6]（semantic segmentation）是计算机视觉领域非常重要的基本任务，其目标是**对图像中的每一个像素进行分类**。使用深度学习方法求解该任务往往需要数据集图片具备非常稠密的像素级别类别标注。
- 标注成本：平均每张图90分钟[14]，2万张图的中等规模数据集需要3万小时。

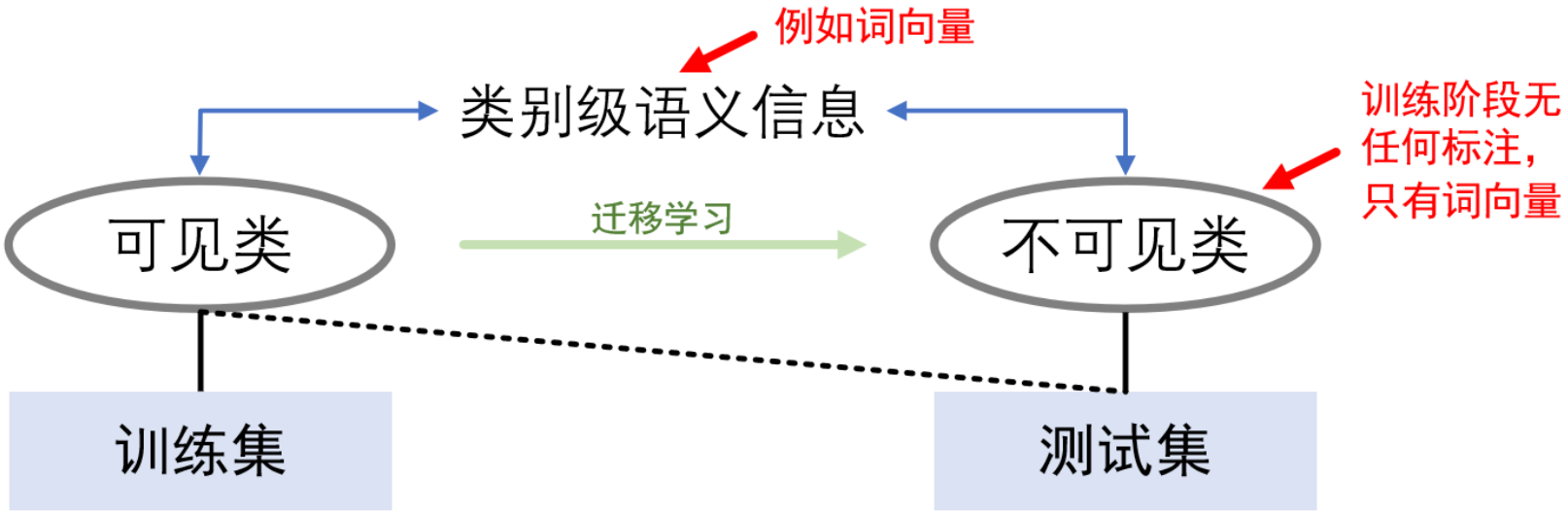
图像 Image

标注 GT mask

研究背景

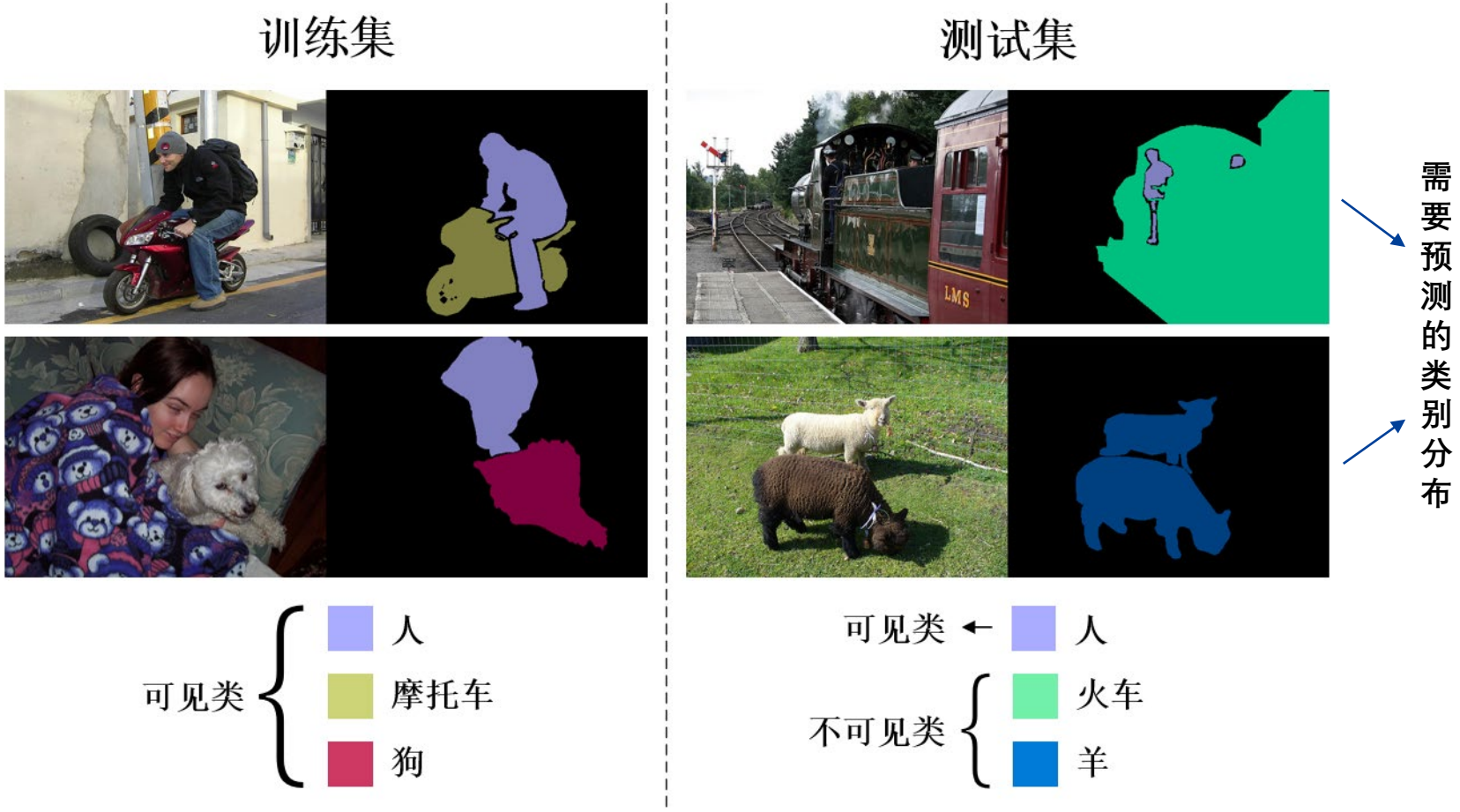
➢ 零示例语义分割

- 为了**减少标注数据集所需的人力以及模型对标注的依赖**，本论文探讨了难度极高的零示例语义分割[7]（zero-shot semantic segmentation）。
- 类别集合被划分为**可见类**（Seen）和**不可见类**（Unseen）。



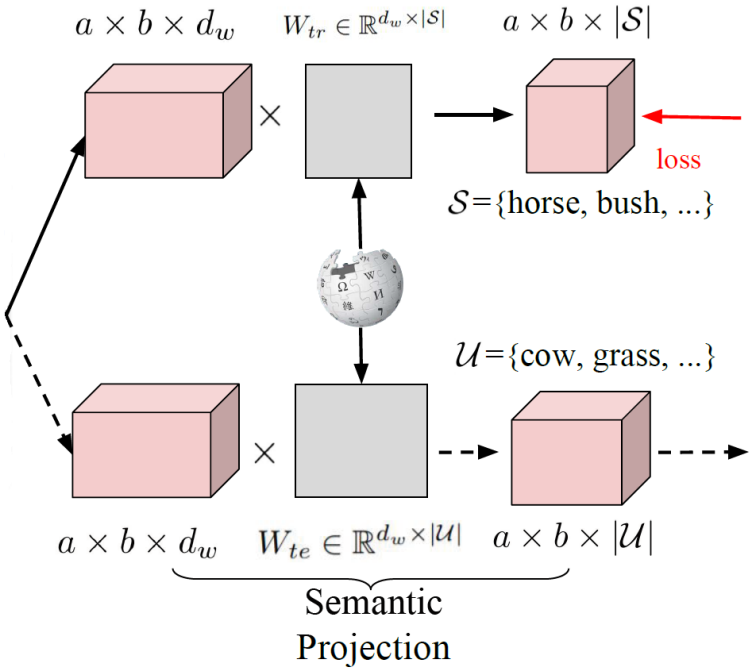
研究背景

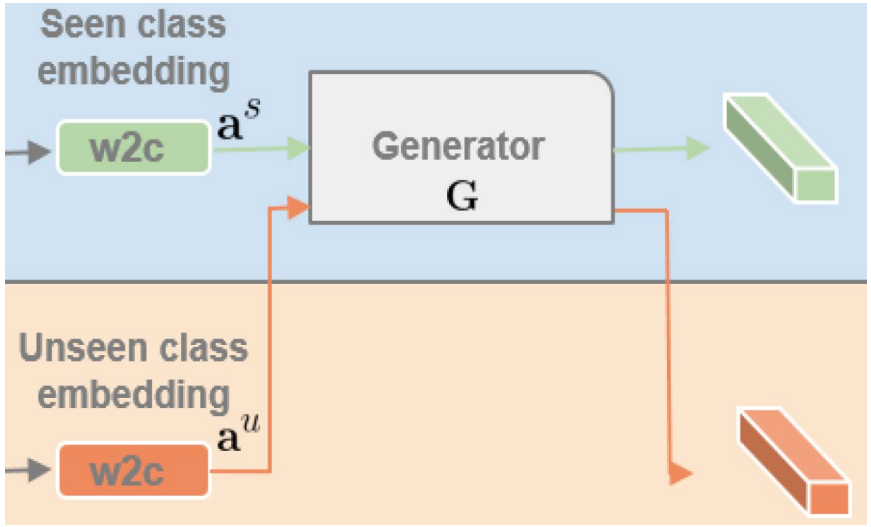
- 一个简单的零示例语义分割数据集



研究背景

- **零示例语义分割任务的当前最优模型（state-of-the-art）**
- **SPNet[8]**（左图）构建了视觉特征空间与词向量空间之间的映射关系。
- **ZS3Net[7]**（右图）使用词向量生成了伪视觉特征。





目录

研究任务

➢ 研究目标

- 对ZS3Net[7]中基于词向量的特征生成方法（feature generation）做出改进，进一步提升零示例语义分割任务的最优结果，着重增强模型对不可见类的分割能力。

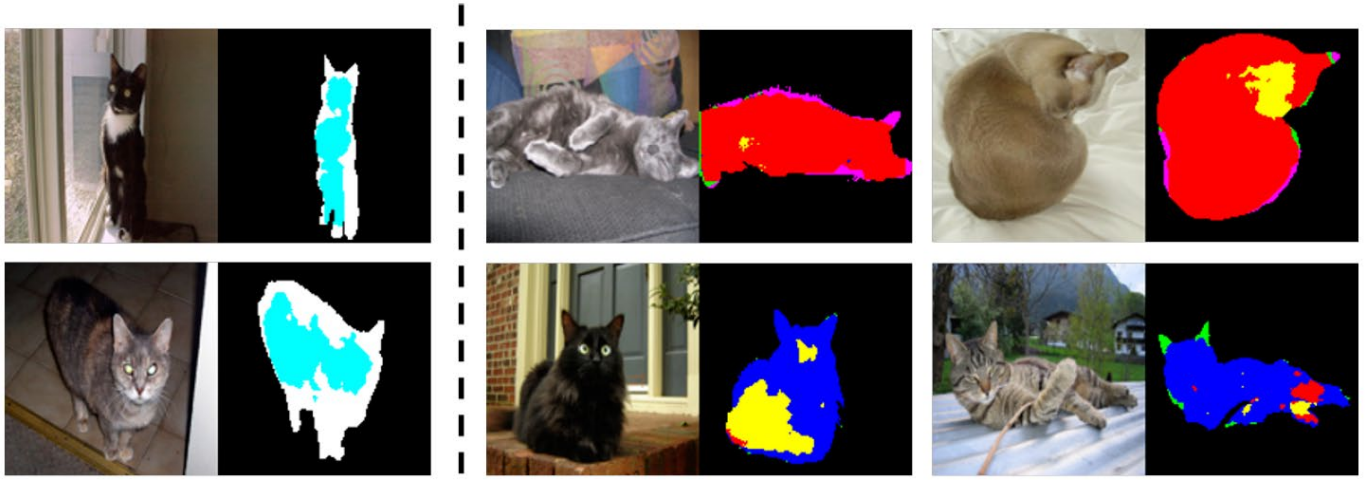
➢ 需要解决的主要问题

- 设计**有效的模型结构**得到多元且真实的伪视觉特征。
- 构建**合理的模型优化策略**达到最优的零示例语义分割结果。

目录

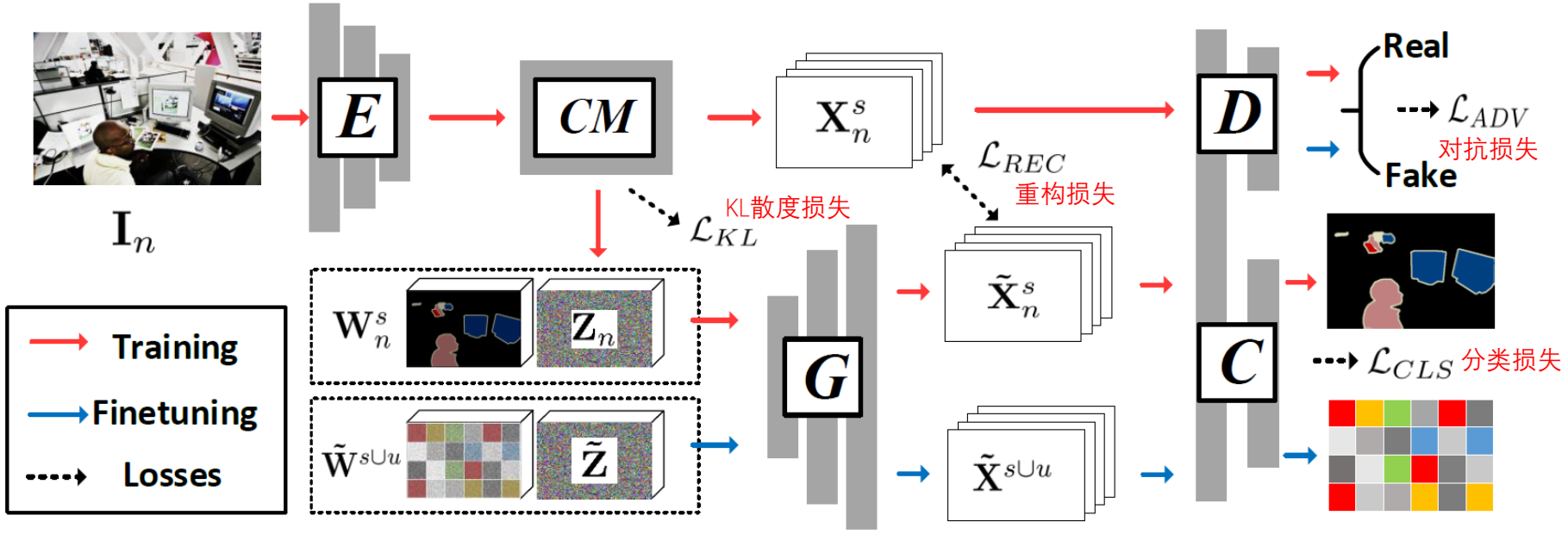
思路来源

- 在视觉特征图上，某个像素级别特征的像素级别上下文信息指的是从与该像素空间相邻的其它像素级别特征中提取的综合信息。
- 我们在 Pascal-Context[9] 数据集上训练了 Deeplabv2[13] 模型，提取出关于类别“猫”的像素级别特征，并用 K 均值聚类算法（K-means）将这些特征像素聚为 K 类。可以看出，像素级别特征往往会非常依赖于像素级别上下文信息。



基本框架和流程

- 本文提出了基于上下文感知生成特征的零示例语义分割模型，它包含骨干网络 E、上下文模块 CM、特征生成器 G、判别器 D 和分类器 C。其中 E 和 CM 提取图像的**真实特征**，CM 引导 G 重构生成**伪造特征**。
- 网络优化分为**训练**和**微调**两阶段。模型在训练阶段学习可见类知识，在微调阶段完成从可见类向不可见类的知识迁移。



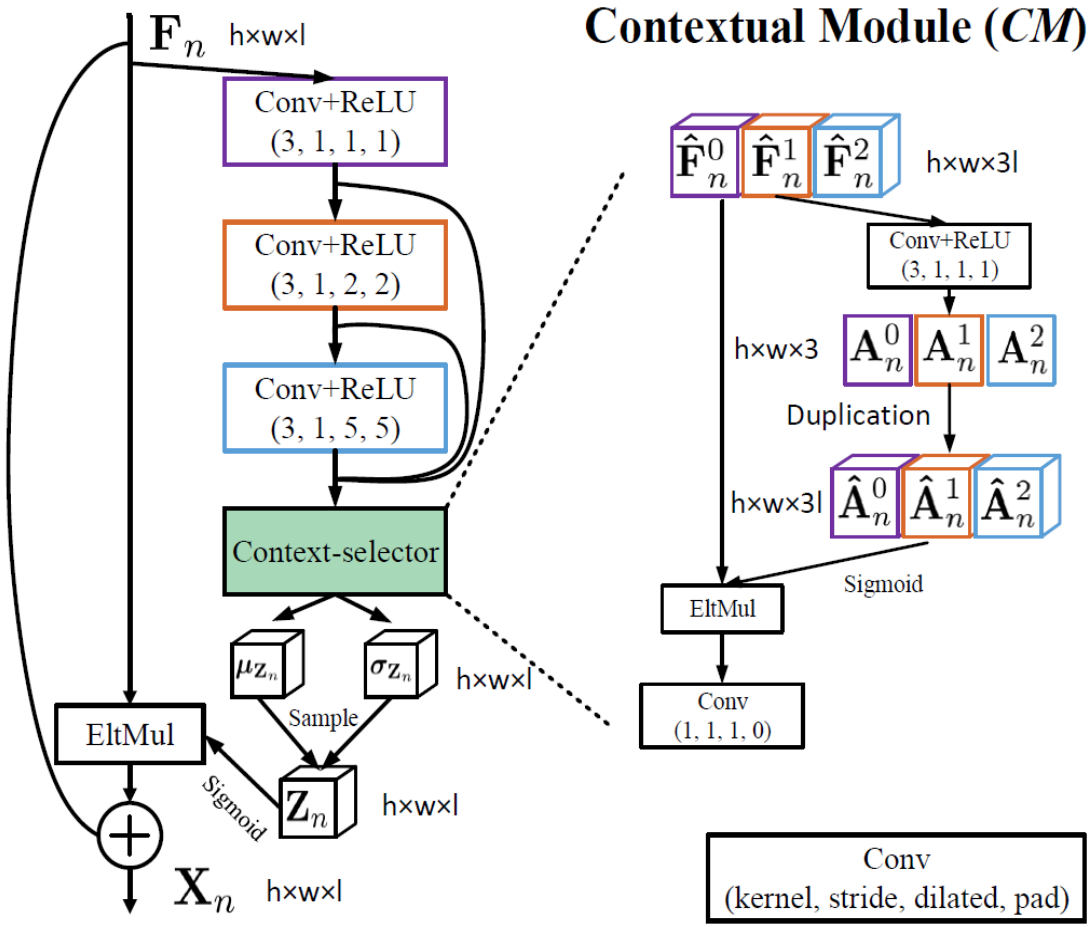
上下文模块

主要作用

- 输出像素级别上下文信息来引导特征生成器生成基于上下文感知的伪造特征。

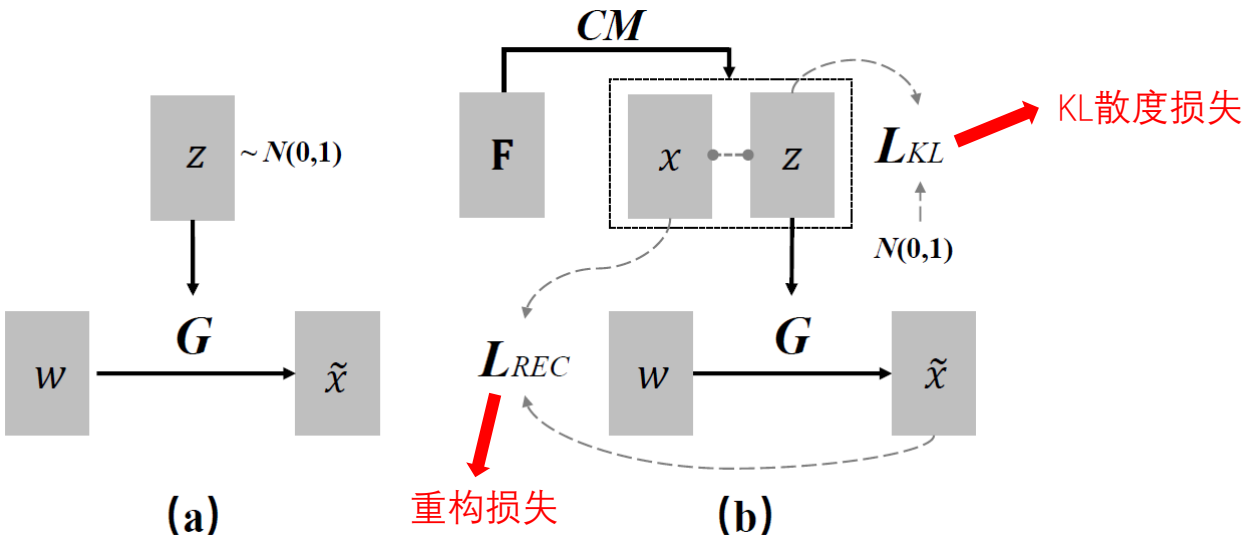
主要结构

- 空洞卷积 Dilated Conv
- 多尺度上下文信息图 $\hat{F}^k$
- 上下文选择器 CS
- 上下文潜在变量 Z
- 原特征图 F -> 新特征图 X



训练阶段本文模型与ZS3Net的对比

- 如图 (a) 所示，ZS3Net 使用词向量 w 和随机噪声 z 作为生成器 G 的输入得到像素级别伪造特征 $\tilde{x}$ 。如图 (b) 所示，本文则使用 w 和**上下文潜在变量** z 作为 G 的输入重构得到 $\tilde{x}$ 。其中上下文潜在变量 z 由上下文模块 CM 输出得到，它编码了**像素级别上下文信息**并支持随机采样。
- 同一个词向量搭配不同上下文潜在变量可以得到**不同上下文环境下的伪造特征**，因此我们**构建了上下文潜在变量与像素级别伪造特征的一一映射关系**。



微调阶段

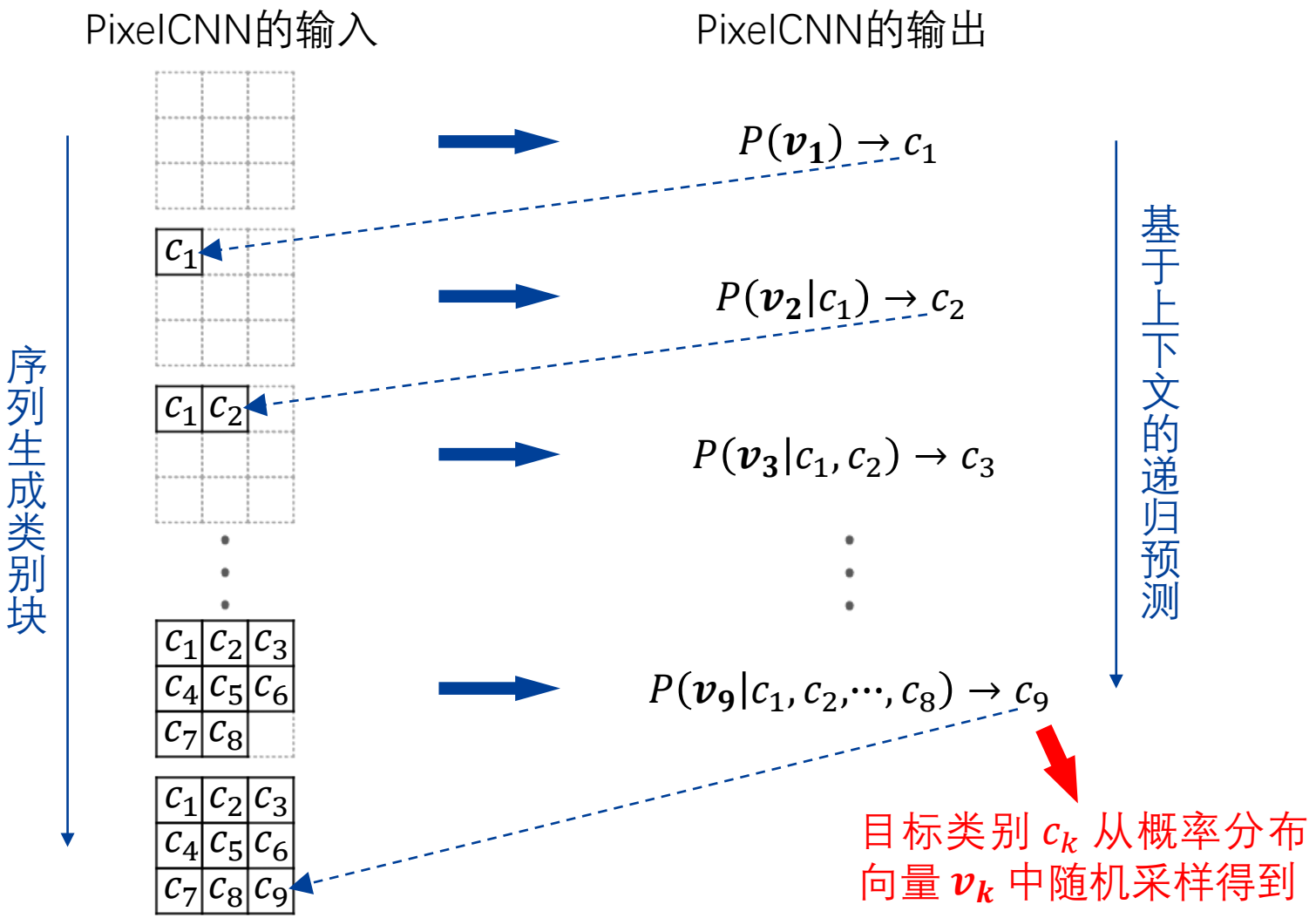
➢ 方案一：基于点的微调（Pixel-wise Finetuning，简称PF）

- 将生成器根据词向量（包含不可见类）和随机噪声生成的像素级别（点状）伪造特征随机堆叠成伪造特征图用于网络优化。
- 伪造特征图相较于真实特征图**缺少了相邻像素之间的关联信息**。我们提出了方案二来解决这一问题。

➢ 方案二：基于块的微调（Block-wise Finetuning，简称BF）

- 使用PixelCNN[12]模型生成由类别像素点（包含不可见类）排列而成的类别块。块内像素点的排列方式更符合真实图片像素点之间的依赖关系。
- 基于类别块的微调**考虑了相邻像素之间的关联信息**，因此类别块引导生成器输出的伪造特征（包含不可见类）更接近真实特征。

基于块的微调中3×3类别块的生成



目录

数据集和评价标准

数据集

- Pascal-Context [9] -> 33类（包含4类不可见类）中等规模数据集
- COCO-stuff [10] -> 182类（包含15类不可见类）大规模数据集
- Pascal-VOC 2012 [11] -> 20类（包含5类不可见类）小规模数据集

评价标准

- harmonic IoU [8]（简称hIoU）-> **最主要指标**
- mean IoU（简称mIoU）-> 重要指标
- pixel accuracy（简称PA）
- mean accuracy（简称MA）

实验结果（定量分析）

- 本文方法Ours(PF)、Ours(BF)与基准方法在三个数据集上的语义分割对比。ST表示自学习策略[7] self-training。
- 本文方法在不可见类评价和综合评价上取得大幅提升。
- 对于hIoU标准，Ours(BF)相较于基准方法在三个数据集上分别带来了67.7%、30.4%和 50.5% 的相对提升。

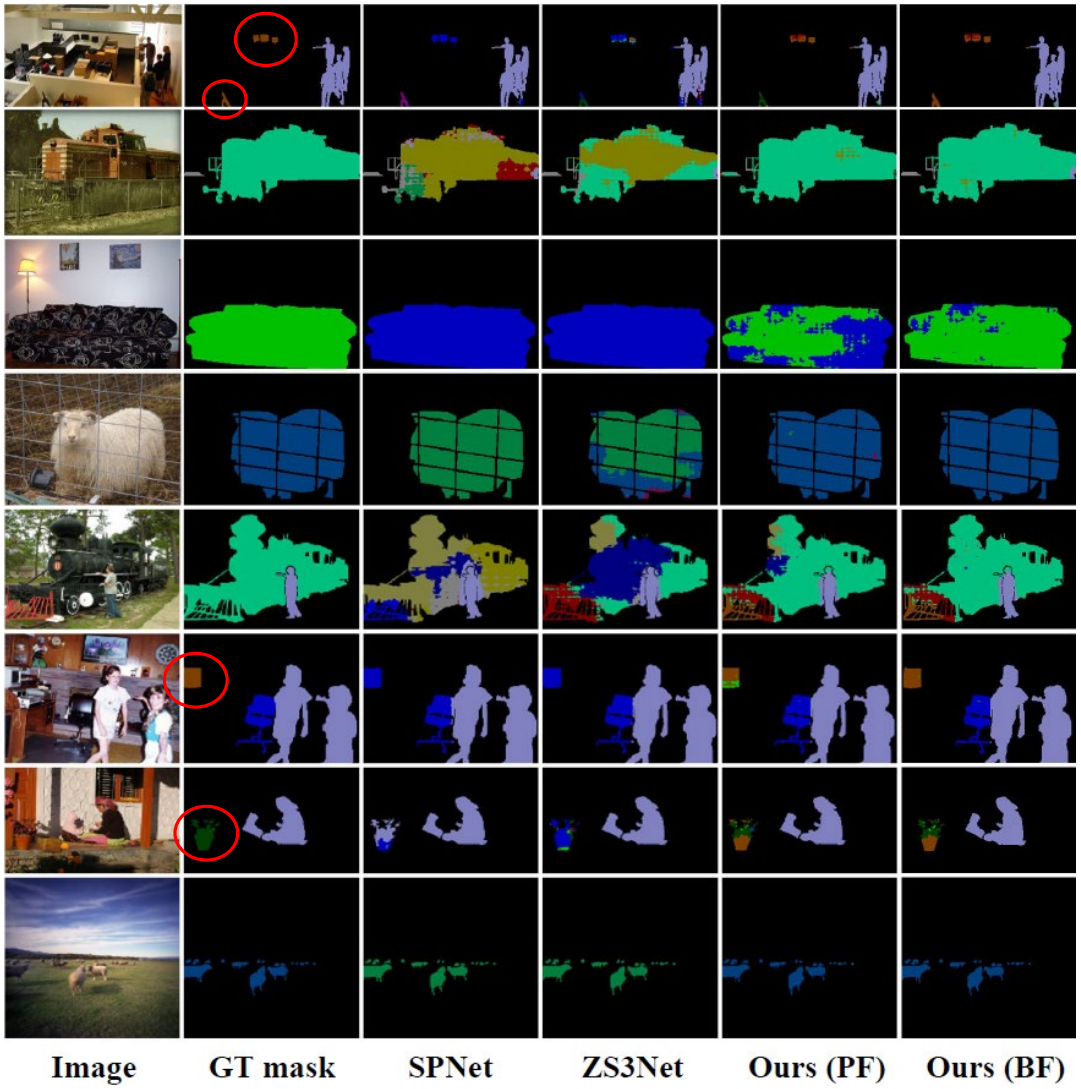
Method	Overall				Seen			Unseen		
	hIoU	mIoU	PA	MA	mIoU	PA	MA	mIoU	PA	MA
Pascal-Context										
SPNet	0	0.2938	0.5793	0.4486	0.3357	**0.6389**	0.5105	0	0	0
SPNet-c	0.0718	0.3079	0.5790	0.4488	0.3514	0.6213	0.4915	0.0400	0.1673	0.1361
ZS3Net	0.1246	0.3010	0.5710	0.4442	0.3304	0.6099	0.4843	0.0768	0.1922	0.1532
Ours(PF)	0.2061	0.3347	**0.5975**	0.4900	0.3610	0.6180	0.5140	0.1442	0.3976	0.3248
Ours(BF)	**0.2089**	**0.3443**	0.5926	**0.5082**	**0.3718**	0.6107	**0.5285**	**0.1453**	**0.4161**	**0.3612**
ZS3Net+ST	0.1488	0.3102	0.5725	0.4532	0.3398	0.6107	0.4935	0.0953	0.2030	0.1721
Ours(PF)+ST	**0.2252**	**0.3352**	**0.5961**	**0.4962**	**0.3644**	**0.6120**	**0.5065**	**0.1630**	**0.5038**	**0.4214**
COCO-stuff										
SPNet	0.0140	0.3164	0.5132	0.4593	0.3461	**0.6564**	0.5030	0.0070	0.0171	0.0007
SPNet-c	0.1398	0.3278	0.5341	0.4363	0.3518	0.6176	0.4628	0.0873	0.2450	0.1614
ZS3Net	0.1495	0.3328	0.5467	0.4837	0.3466	0.6434	0.5037	0.0953	0.2275	0.2701
Ours(PF)	0.1819	**0.3345**	**0.5658**	0.4845	**0.3549**	0.6562	0.5066	0.1223	0.2545	0.2701
Ours(BF)	**0.1949**	0.3054	0.5476	**0.4918**	0.3202	0.6198	**0.5086**	**0.1401**	**0.2988**	**0.3177**
ZS3Net+ST	0.1620	0.3367	0.5631	**0.4862**	0.3489	0.6584	0.5042	0.1055	0.2488	0.2718
Ours(PF)+ST	**0.1946**	**0.3372**	**0.5676**	0.4854	**0.3555**	**0.6587**	**0.5058**	**0.1340**	**0.2670**	**0.2728**
Pascal-VOC										
SPNet	0.0002	0.5687	0.7685	0.7093	0.7583	**0.9482**	**0.9458**	0.0001	0.0007	0.0001
SPNet-c	0.2610	0.6315	0.7755	0.7188	0.7800	0.8877	0.8791	0.1563	0.2955	0.2387
ZS3Net	0.2874	0.6164	0.7941	0.7349	0.7730	0.9296	0.8772	0.1765	0.2147	0.1580
Ours(PF)	0.3972	0.6545	**0.8068**	0.7636	**0.7840**	0.8950	0.8868	0.2659	0.4297	0.3940
Ours(BF)	**0.4326**	**0.6623**	**0.8068**	**0.7643**	0.7833	0.8745	0.8621	**0.2988**	**0.5176**	**0.4710**
ZS3Net+ST	0.3328	0.6302	0.8095	0.7382	0.7802	**0.9189**	**0.8569**	0.2115	0.3407	0.2637
Ours(PF)+ST	**0.4366**	**0.6577**	**0.8164**	**0.7560**	**0.7859**	0.8704	0.8390	**0.3031**	**0.5855**	**0.5071**

实验结果（定性分析）

- 基准方法和本文方法在Pascal-VOC数据集上的零示例语义分割可视化结果。

- 本文方法能够更有效地分割不可见类物体：

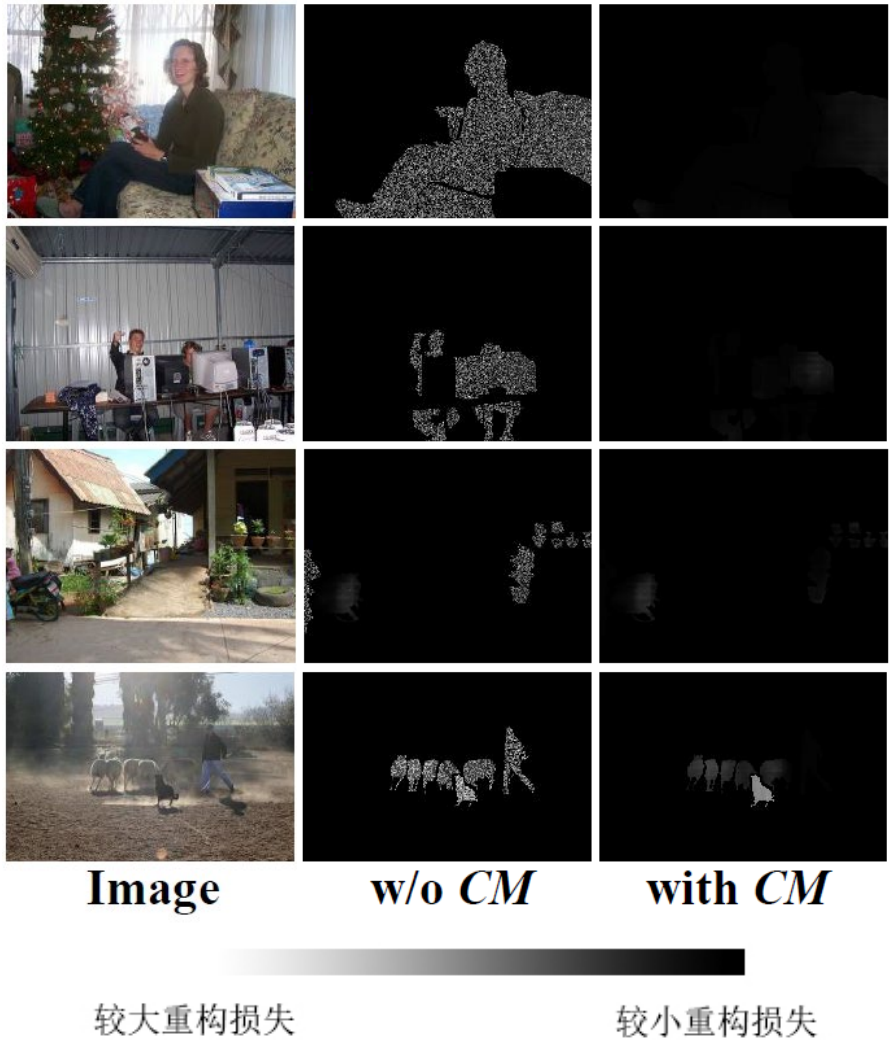
  显示器（第一、六行，棕色）
  火车（第二、五行，淡绿色）
  沙发（第三行，翠绿色）
  绵羊（第四、八行，深蓝色）
  盆栽植物（第七行，深绿色）



实验结果（定性分析）

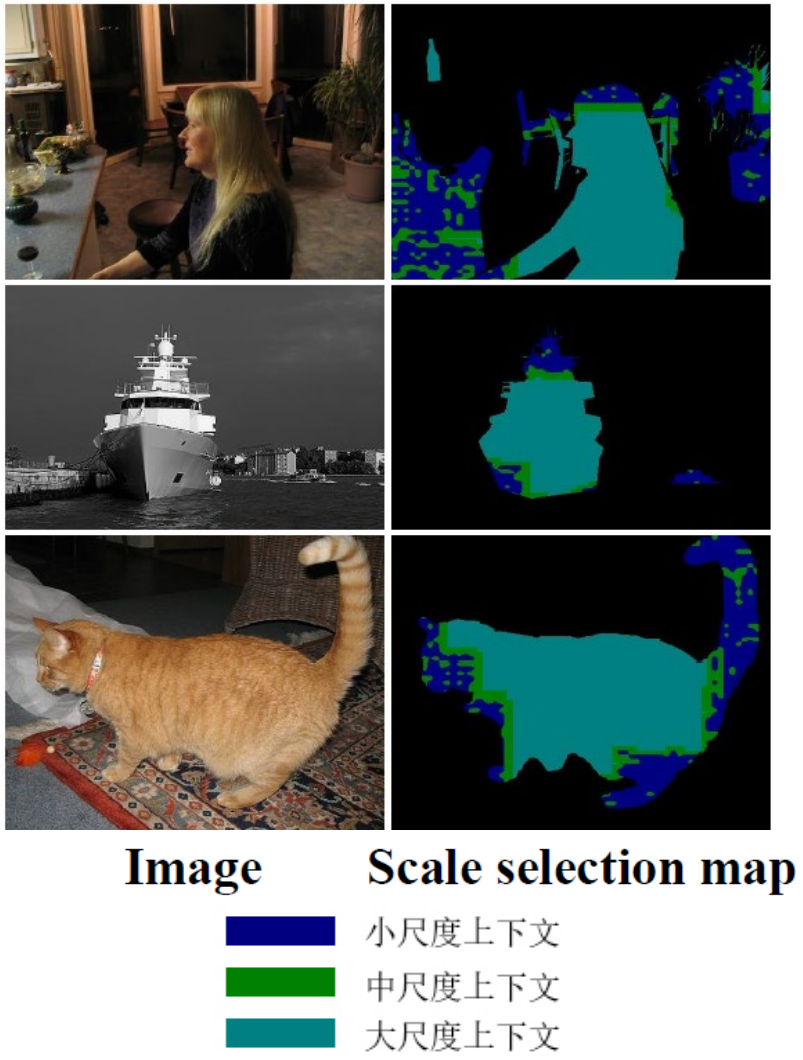
- 基于Pascal-VOC数据集测试集图片的特征重构质量可视化结果。
- 右图展示了真实特征图和伪造特征图之间的重构损失分布。
- 加入 CM 能准确重构出可见类甚至不可见类的像素级别特征：

  沙发（第一行）
  显示器（第二行）
  盆栽植物（第三行）
  绵羊（第四行）



实验结果（定性分析）

- 基于Pascal-VOC数据集的上下文选择器 CS 可视化效果。
- CS 可以推断出影响每个像素级别特征的上下文范围大小（小尺度、中尺度或大尺度）。
- 具有区分性的局部区域内的像素点趋于受到小尺度上下文信息的影响。其余像素点趋于受到中尺度或大尺度上下文信息的影响。



目录

主要贡献

- 创造性地提出了**基于上下文感知的特征生成方法**，利用生成器得到了更加多元且真实的伪造特征，大幅提升了零示例语义分割任务的最优结果。
- 三点细化贡献：1）成功将语义分割网络和特征生成网络进行联合；2）设计了带有新颖上下文选择器的上下文模块；3）在优化阶段提出了基于块的微调方案，进一步提升了语义分割模型的迁移能力。
- 在三个标准语义分割数据集上进行的大量定量/定性实验都证明本文提出的方法具备目前**最优的分割性能**以及基准方法不具备的**延展功能**。

参考文献

[1] LONG J, SHELHAMER E, DARRELL T. Fully Convolutional Networks For Semantic Segmentation[C]. CVPR, 2015.

[2] ZHAO H, SHI J, QI X, et al. Pyramid Scene Parsing Network[C]. CVPR, 2017.

[3] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, And Fully Connected CRFs[J]. TPAMI, 2018, 40(4): 834-848.

[4] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks For Biomedical Image Segmentation[C]. MICCAI, 2015.

[5] LIN G, MILAN A, SHEN C, et al. RefineNet: Multi-Path Refinement Networks For High-Resolution Semantic Segmentation[C]. CVPR, 2017.

[6] ZHANG Z, CHEN A, XIE L, et al. Learning Semantics-Aware Distance Map With Semantics Layering Network For Amodal Instance Segmentation[C]. ACMMM, 2019.

[7] BUCHER M, VU T H, CORD M, et al. Zero-Shot Semantic Segmentation[C]. NeurIPS, 2019.

[8] XIAN Y, CHOUDHURY S, HE Y, et al. Semantic Projection Network For Zero-and Few-Label Semantic Segmentation[C]. CVPR, 2019.

[9] MOTTAGHI R, CHEN X, LIU X, et al. The Role Of Context For Object Detection And Semantic Segmentation In The Wild[C]. CVPR, 2014.

参考文献

[10] CAESAR H, UIJLINGS J, FERRARI V. Coco-Stuff: Thing And Stuff Classes In Context[C]. CVPR, 2018.

[11] EVERINGHAM M, ESLAMI S M A, VAN GOOL L, et al. The Pascal Visual Object Classes Challenge: A Retrospective[J]. IJCV, 2015, 111(1): 98-136.

[12] OORD A V D, KALCHBRENNER N, KAVUKCUOGLU K. Pixel Recurrent Neural Networks[C]. ICML, 2016.

[13] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, And Fully Connected CRFs[J]. TPAMI, 2018, 40(4): 834-848.

[14] Cordts M, Omran M, Ramos S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C]. CVPR, 2016.