



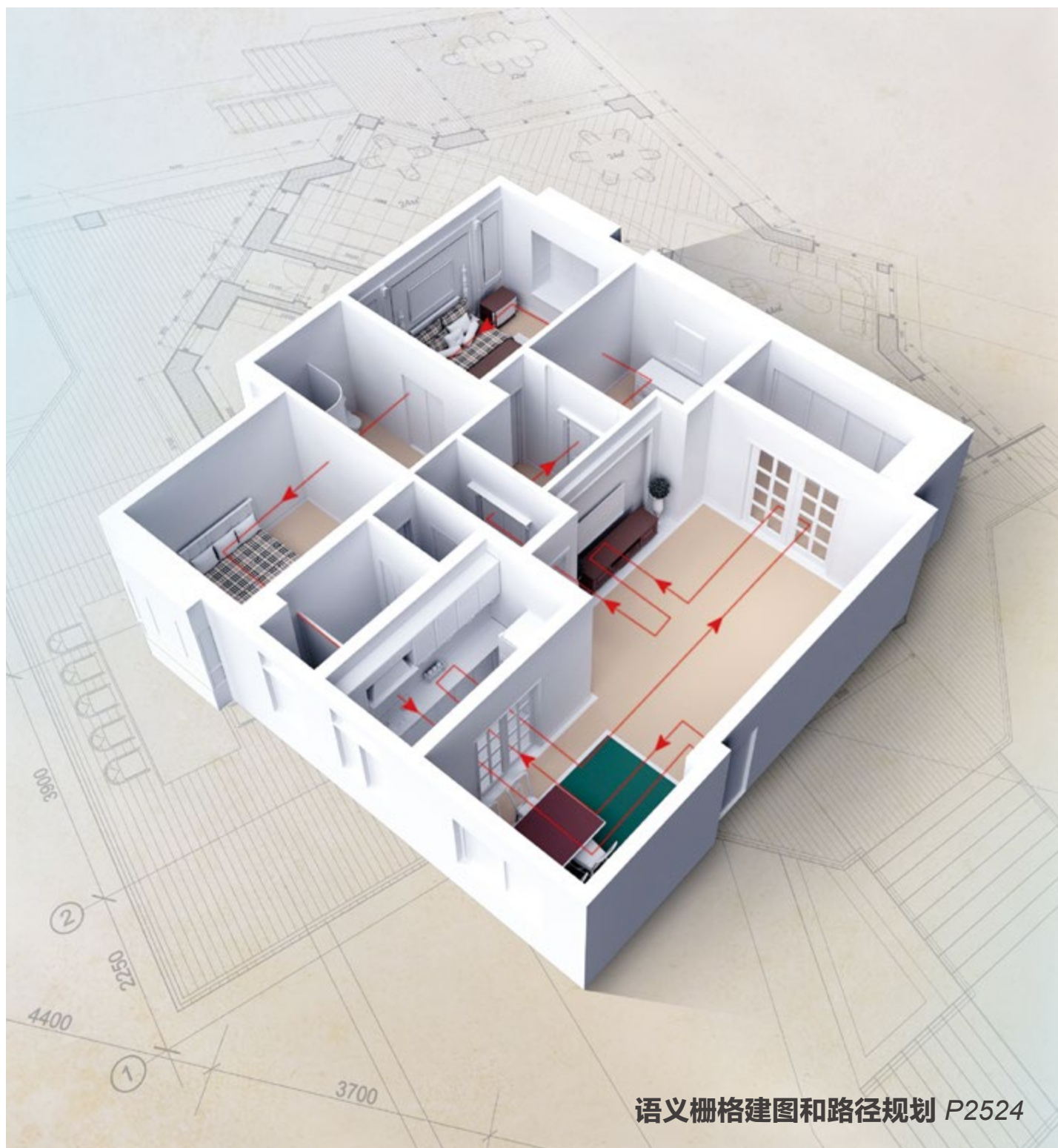
JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
10
VOL.26

ISSN1006-8961
CN11-3758/TB



语义栅格建图和路径规划 P2524



第26卷第10期（总第306期）
2021年10月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主 编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮 编 100190
电子信箱 jig@aircas.ac.cn
电 话 010-58887035
网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订 购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
(邮政信箱: 北京399信箱 邮编: 100048)
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian

Monthly, Started in 1996

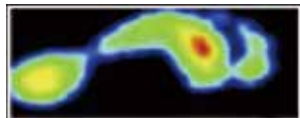
Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元



穿鞋足迹序列的足迹能量图组表达与识别(第2357页)



时空图卷积网络与注意机制的视频目标分割(第2376页)



双光流网络指导的视频目标检测(第2473页)

综述

图像分类的深度卷积神经网络模型综述

张珂, 冯晓晗, 郭玉荣, 苏昱坤, 赵凯, 赵振兵, 马占宇, 丁巧林 2305

中国图像工程25年

章毓晋 2326

数据集论文

FGSC-23: 面向深度学习精细识别的高分辨率光学遥感图像舰船目标数据集

姚力波, 张筱晗, 吕亚飞, 孙炜玮, 李孟洋 2337

图像处理和编码

并行生成网络的红外—可见光图像转换

余佩伦, 施佳, 王晗 2346

图像分析和识别

穿鞋足迹序列的足迹能量图组表达与识别

王新年, 于丹, 张涛 2357

时空图卷积网络与注意机制的视频目标分割

姚睿, 夏士雄, 周勇, 赵佳琦, 胡伏原 2376

多支路协同的RGB-T图像显著性目标检测

蒋亭亭, 刘昱, 马欣, 孙景林 2388

改进R-FCN模型的小尺度行人检测

刘万军, 董利兵, 曲海成 2400

边缘与区域不一致性引导下的图像拼接篡改检测网络

蒋小玉, 刘春晓 2411

图像理解和计算机视觉

各向异性导向滤波的红外与可见光图像融合

刘明威, 王任华, 李静, 焦映臻 2421

卷积稀疏与细节显著图解析的图像融合

杨培, 高雷阜, 訾玲玲 2433

计算机图形学

利用距离与内能极小平滑链接Bézier曲线

李军成, 刘成志, 刘珊君 2450

NCIG 2020

结合置信度加权融合与视觉注意机制的前景检测

成科扬, 孙爽, 王文杉, 师文喜, 李鹏, 詹永照 2462

双光流网络指导的视频目标检测

尉婉青, 禹晶, 史薪琪, 肖创柏 2473

生成对抗网络驱动的图像隐写与水印模型

郑钢, 胡东辉, 戈辉, 郑淑丽 2485

融合语义—表观特征的无监督前景分割

李熹, 马惠敏, 马洪兵, 王弈冬 2503

场景视点偏移的激光雷达点云分割

郑阳, 林春雨, 廖康, 赵耀, 薛松 2514

激光—相机系统语义栅格建图和路径规划

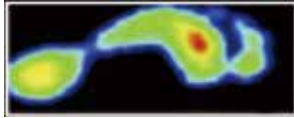
丁梦远, 郭迟, 黄凯 2524

融合词袋连通图的图像检索特征选择

李国祥, 王继军, 马文斌 2533

CONTENTS

JOURNAL OF IMAGE AND GRAPHICS



Shoeprint sequence representation and recognition using shoeprint energy map set (P2357)



Spatial-temporal video object segmentation with graph convolutional network and attention mechanism(P2376)



Dual optical flow network-guided video object detection(P2473)

Review

Overview of deep convolutional neural networks for image classification

Zhang Ke, Feng Xiaohan, Guo Yurong, Su Yukun, Zhao Kai, Zhao Zhenbing, Ma Zhanyu, Ding Qiaolin 2305

Twenty-five years of image engineering in China

Zhang Yujin 2326

Dataset

FGSC-23: a large-scale dataset of high-resolution optical remote sensing image for deep learning-based fine-grained ship recognition

Yao Libo, Zhang Xiaohan, Lyu Yafei, Sun Weiwei, Li Mengyang 2337

Image Processing and Coding

Infrared-to-visible image translation based on parallel generator network

Yu Peilun, Shi Quan, Wang Han 2346

Image Analysis and Recognition

Shoeprint sequence representation and recognition using shoeprint energy map set

Wang Xinnian, Yu Dan, Zhang Tao 2357

Spatial-temporal video object segmentation with graph convolutional network and attention mechanism

Yao Rui, Xia Shixiong, Zhou Yong, Zhao Jiaqi, Hu Fuyuan 2376

Multi-path collaborative salient object detection based on RGB-T images

Jiang Tingting, Liu Yu, Ma Xin, Sun Jinglin 2388

Small-scale pedestrian detection based on improved R-FCN model

Liu Wanjun, Dong Libing, Qu Haicheng 2400

Edge and region inconsistency-guided image splicing tamper detection network

Jiang Xiaoyu, Liu Chunxiao 2411

Image Understanding and Computer Vision

Infrared and visible image fusion with multi-scale anisotropic guided filtering

Liu Mingwei, Wang Renhua, Li Jing, Jiao Yingzhen 2421

Image fusion method of convolution sparsity and detail saliency map analysis

Yang Pei, Gao Leifu, Zi Lingling 2433

Computer Graphics

Smoothing linked Bézier curves by distance and internal energy minimization

Li Juncheng, Liu Chengzhi, Liu Shanjun 2450

NCIG 2020

Foreground detection via fusing confidence by weight and visual attention mechanism

Cheng Keyang, Sun Shuang, Wang Wenshan, Shi Wenxi, Li Peng, Zhan Yongzhao 2462

Dual optical flow network-guided video object detection

Yu Wanjing, Yu Jing, Shi Xinqi, Xiao Chuangbai 2473

End-to-end image steganography and watermarking driven by generative adversarial networks

Zheng Gang, Hu Donghui, Ge Hui, Zheng Shuli 2485

Semantic-apparent feature-fusion-based unsupervised foreground segmentation method

Li Xi, Ma Huimin, Ma Hongbing, Wang Yidong 2503

LiDAR point cloud segmentation through scene viewpoint offset

Zheng Yang, Lin Chunyu, Liao Kang, Zhao Yao, Xue Song 2514

Semantic grid mapping and path planning combined with laser-camera system

Ding Mengyuan, Guo Chi, Huang Kai 2524

Feature selection method for image retrieval based on connected graphs and bag of words

Li Guoxiang, Wang Jijun, Ma Wenbin 2533

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2021)10-2503-11

论文引用格式: Li X, Ma H M, Ma H B and Wang Y D. 2021. Semantic-apparent feature-fusion-based unsupervised foreground segmentation method. Journal of Image and Graphics, 26(10): 2503-2513 (李熹, 马惠敏, 马洪兵, 王弈冬. 2021. 融合语义—表观特征的无监督前景分割. 中国图象图形学报, 26(10): 2503-2513) [DOI: 10.11834/jig.200442]

融合语义—表观特征的无监督前景分割

李熹¹, 马惠敏^{2*}, 马洪兵^{1,3}, 王弈冬¹

1. 清华大学, 北京 100084; 2. 北京科技大学, 北京 100083; 3. 新疆大学, 乌鲁木齐 830046

摘要: 目的 前景分割是图像理解领域中的重要任务, 在无监督条件下, 由于不同图像、不同实例往往具有多变的表达形式, 这使得基于固定规则、单一类型特征的方法很难保证稳定的分割性能。针对这一问题, 本文提出了一种基于语义—表观特征融合的无监督前景分割方法 (semantic apparent feature fusion, SAFF)。方法 基于语义特征能够对前景物体关键区域产生精准的响应, 但往往产生的前景分割结果只关注于关键区域, 缺乏物体的完整表达; 而以显著性、边缘为代表的表观特征则提供了更丰富的细节表达信息, 但基于表观规则无法应对不同的实例和图像成像模式。为了融合表观特征和语义特征优势, 研究建立了融合语义、表观信息的一元区域特征和二元上下文特征编码的方法, 实现了对两种特征表达的全面描述。接着, 设计了一种图内自适应参数学习的方法, 用于计算最适合的特征权重, 并生成前景置信分数图。进一步地, 使用分割网络来学习不同实例间前景的共性特征。结果 通过融合语义和表观特征并采用图像间共性语义学习的方法, 本文方法在 PASCAL VOC (pattern analysis, statistical modelling and computational learning visual object classes) 2012 训练集和验证集上取得了显著超过类别激活映射 (class activation mapping, CAM) 和判别性区域特征融合方法 (discriminative regional feature integration, DRFI) 的前景分割性能, 在 F 测度指标上分别提升了 3.5% 和 3.4%。结论 本文方法可以将任意一种语义特征和表观特征前景计算模块作为基础单元, 实现对两种策略的融合优化, 取得了更优的前景分割性能。

关键词: 计算机视觉; 前景分割; 无监督学习; 语义—表观特征融合; 自然场景图像; PASCAL VOC 数据集; 自适应加权

Semantic-apparent feature-fusion-based unsupervised foreground segmentation method

Li Xi¹, Ma Huimin^{2*}, Ma Hongbing^{1,3}, Wang Yidong¹

1. Tsinghua University, Beijing 100084, China; 2. University of Science and Technology Beijing, Beijing 100083, China;
3. Xinjiang University, Urumqi 830046, China

Abstract: **Objective** Foreground segmentation is an essential research in the field of image understanding, which is a pre-processing step for saliency object detection, semantic segmentation, and various pixel-level learning tasks. Given an image, this task aims to provide each pixel a foreground or background annotation. For fully supervision-based methods, satisfactory results can be achieved via multi-instance-based learning. However, when facing the problem under unsupervised conditions, achieving a stable segmentation performance based on fixed rules or a single type of feature is difficult

收稿日期: 2020-08-21; 修回日期: 2020-11-19; 预印本日期: 2020-11-26

* 通信作者: 马惠敏 mhmpub@ustb.edu.cn

基金项目: 国家自然科学基金项目 (U20B2062; 61773231); 国家重点研发计划项目 (2016YFB0100901); 北京市科学技术项目 (Z191100007419001)

Supported by: National Natural Science Foundation of China (U20B2062; 61773231); National Key Research and Development Program of China (2016YFB0100901); Beijing Municipal Science & Technology Project (Z191100007419001)

because different images and instances always have variable expressions. Moreover, we find that different types of method have different advantages and disadvantages on different aspects. On the one hand, semantic feature-based learning methods could provide accurate key region extraction of foregrounds but could not generate complete object region and edges in detail. On the other hand, richer detailed expression can be obtained based on an apparent feature-based framework, but it cannot be suitable for variable kinds of cases.

Method Based on the observations, we propose an unsupervised foreground segmentation method based on semantic-apparent feature fusion. First, given a sample, we encode it as semantic and apparent feature map. We use a class activation mapping model pretrained on ImageNet for semantic heat map generation and select saliency and edge maps to express the apparent feature. Each kind of semantic and apparent feature can be used, and the established framework is widely adaptive for each case. Second, to combine the advantages of the two type of features, we split the image as super pixels, and set the expression of four elements as unary and binary semantic and apparent feature, which realizes a comprehensive description of the two types of expressions. Specifically, we build two binary relation matrices to measure the similarity of each pair of super pixels, which are based on apparent and semantic feature. For generating the binary semantic feature, we use the apparent feature-based similarity measure as a weight to provide the element for each super pixel, in which semantic-feature-based similarity measure is utilized for binary apparent feature calculation. Based on the different view for feature encoding, the two types of information could be fused for the first time. Then, we propose a method for adaptive parameter learning to calculate the most suitable feature weights and generate the foreground confidence score map. Based on the four elements, we could establish an equation to express each super pixel's foreground confidence score using the least squares method. For an image, we first select super pixels with higher confident scores of unary semantic and apparent feature on foreground or background. Then, we can learn weights of the four elements and bias' linear combination by least squares estimation. Based on the adaptive parameters, we can achieve a better confidence score inference for each super pixel individually. Third, we use segmentation network to learn foreground common features from different instances. In a weakly supervised semantic segmentation task, the fully supervision-based framework is used for improving pseudo annotations for training data and providing inference results. Inspired by the idea, we use the convolution network to mine foreground common feature from different instances. The trained model could be utilized to optimize the quality of foreground segmentation for both images used for network training and new data directly. A better performance can be achieved by fusing semantic and apparent features as well as cascading the modules of intra image adaptive feature weight learning and inter-image common feature learning.

Result We test our methods on the pattern analysis, statistical modelling and computational learning visual object classes (PASCAL VOC) 2012 training and evaluation set, which include 10 582 and 1 449 samples, respectively. Precision-recall curve as well as F-measure are used as indicators to evaluate the experimental results. Compared with typical semantic and apparent feature-based foreground segmentation methods, the proposed framework achieves superior improvement of baselines. For PASCAL VOC 2012 training set, the F-measure has a 3.5% improvement, while a 3.4% increase is obtained on the validation set. We also focus on the performance on visualized results for analysis the advantages of fusion framework. Based on comparison, we can find that results with accurate, detailed expression can be achieved based on the adaptive feature fusion operation, while incorrect cases can further be modified via multi-instance-based learning framework.

Conclusion In this study, we propose a semantic-apparent feature fusion method for unsupervised foreground segmentation. Given an image as input, we first calculate the semantic and apparent feature of the unary region of each super pixel in image. Then, we integrate two types of features through the cross-use of similarity measure of apparent and semantic feature. Next, we establish a context relationship for each pair of super pixels to calculate the binary feature of each region. Further, we establish an adaptive weight learning strategy. We obtain the weighting parameters for optimal foreground segmentation and achieve the confidence in the image foreground by automatically adjusting the influence of each dimensional feature on the foreground estimation in each specific image instance. Finally, we build a foreground segmentation network model to learn the common features of foreground between different instances and samples. Using the trained network model, the image can be re-inferred to obtain more accurate foreground segmentation results. The experiments on the PASCAL VOC 2012 training set and validation set prove the effectiveness and generalization ability of the algorithm. Moreover, the method proposed can use other foreground segmentation methods as a baseline and is widely used to improve the performance of tasks such as foreground segmentation and

weakly supervised semantic segmentation. We also believe that to consider the introduction of various types of semantic and apparent feature fusion as well as adopt alternate iterations to mine the internal spatial context information of image and the common expression features between different instance is a feasible way to improve the performance of foreground segmentation further and an important idea for semantic segmentation tasks.

Key words: computer vision; foreground segmentation; unsupervised learning; semantic-apparent feature fusion; natural scene images; PASCAL VOC dataset; adaptive weighting

1 研究背景

前景分割是图像目标检测(Li等,2018b;Chen等,2016)、语义分割(Chen等,2017;Wang等,2018c)、显著性检测(Li等,2018c)等领域中的重要任务之一。给定一幅图像,前景分割的任务是提供一种实现像素级前景、背景分类的模型。在现有研究中,基于全监督学习的方法(Chen等,2018;Long等,2015)已经取得了良好的性能。然而,在很多实际情况中,经常面临着精确的像素级标注难以大量、高效获取的问题。因此,在弱监督、无监督条件下实现前景分割和语义分割,成为一项具有重要意义的工作。

本文面向无监督前景分割任务,提出了一种语义—表观特征融合的分割方法(semantic apparent feature fusion,SAFF)。研究发现,图像中的高维语义特征往往能够实现对前景关键区域的捕捉,但缺乏细节信息;而以显著性、边缘等为代表的表观特征,则更好地提供了图像细节表达和对局部区域间相似关系的描述,但缺乏对前景物体的语义表达。同时,对于不同的图像、实例而言,其语义特征和表观特征具有不同的表达能力和精度。因此,本文建立了一种自适应参数学习的方式,提出了一种对每幅图像语义和表观特征的最优加权策略,实现了对前景区域的精确提取。接着,使用多实例学习的框架以实现不同图像前景共性特征的计算,并进一步实现了对图像前景分割结果的重推理和性能优化。

在已有研究中,基于弱监督 and 全监督的语义分割方法为本文提供了参考。一方面,现有面向前景及语义分割任务的研究,提供了多种全监督条件的学习框架。以全卷积网络(fully convolutional network,FCN)(Long等,2015)为基础,DeepLab(Chen等,2018)及其变体通过引入空洞卷积(Chen等,2018;Wang等,2018a)、条件随机场(conditional ran-

dom fields,CRFs)(Zheng等,2015)、深度空间金字塔池化(atrous spatial pyramid pooling,ASPP)等单元,得到了更高的性能。而针对无监督任务,虽然基于全监督的深度学习框架不能直接实现对像素级分割的学习与推理,但其仍然可以用来学习不同图像和实例间的共性表达。在本文中,基于大量图像样本和对应的初始分割结果,运用全监督框架来实现对前景共性特征的学习。进一步地,利用训练好的分割网络,每一幅图像可以被重新推理得到性能更高的分割结果。

另一方面,当前弱监督条件下的前景或语义分割任务得到了广泛的关注。其中,只利用图像标签的弱监督条件最具有难度。在这一设置下,采用两阶段式的方法取得了更高的性能,即首先生成伪标签信息,再利用全监督框架实现语义分割过程(Wang等,2018c;Ahn和Kwak,2018;Huang等,2018;Li等,2020)。在伪标签生成中,类别激活映射(class activation mapping,CAM)模型(Zhou等,2016)被广泛用来计算图像对分类贡献具有高响应的语义区域。基于初始定位种子,AffinityNet(Ahn和Kwak,2018)提出了一种从伪标签中学习像素间相似关系的方法,实现了对初始标注精度的提升;DSRG(deep seed region growing)(Huang等,2018),MCOF(mine common object feature)(Wang等,2018c),WRSS(weaklier supervised segmentation)(Li等,2020)则分别考虑在全监督训练过程中利用不同的迭代结构提升分割的性能。受到这些工作启发,本文同样采用两阶段的方式实现前景分割任务。对于一幅图像而言,由于其只包含前景和背景两个类别,因此无监督前景分割任务可以看为是弱监督语义分割的一种变体。本研究首先建立了基于图内上下文信息的前景学习模型,进一步将生成的特征图作为伪标签信息,利用分割网络实现图间多实例学习过程,最终得到高质量的分割结果。

本文的贡献主要包括以下3个方面:

1) 提出了一种语义和表观特征联合编码的无监督图像前景分割方法,并建立了一种图像内上下文相似度计算模型,得到了对图像前景信息的判别性表达。

2) 提出了一种自适应权重的图像内区域前景置信度学习和推理方法,自适应地实现对不同图像、实例的最优特征加权,实现了高质量的分割结果。分割网络被级联使用来学习多实例间的共性特征,并进一步提升了性能。

3) 针对无监督前景分割任务,本文方法在 PASCAL VOC (pattern analysis, statistical modelling and computational learning visual object classes) 2012 数据集上得到的分割结果取得了明显高于基线的性能。

2 算法框架

给定一幅图像,首先将其编码为语义特征和表观特征。具体地,研究采用在 ImageNet (Deng 等, 2009) 上预训练的特征提取器作为高维语义编码模块,用显著性、边缘特征编码图像的表观信息。同

时,计算图像的超像素分割结果。对于每个超像素,以 4 维特征来描述其前景置信度。具体地,首先利用语义特征各维度的最大响应值和显著性值来分别描述该超像素区域的一元区域语义、表观特征。接着,定义了两两超像素之间的语义特征相似度和边缘连通度,结合图像一元超像素区域特征,计算生成了二元上下文语义和表观特征。语义特征相似度、边缘连通度的计算参考了 Wang 等人(2018b)和 Li 等人(2018a)的工作。

进一步地,提出了一种自适应权重计算的方法。对于每个超像素,以这 4 维特征作为输入,通过自适应学习得到的权重,实现对前景置信度的推理。对于 N 幅图像样本,将得到 N 个对应的前景分割结果。此时,将这些前景分割结果作为伪标签,采用全监督分割网络的框架,实现了对图像多实例间前景共性特征的学习。基于训练完成的模型,不仅这 N 幅用于训练的图像样本可以被重推理优化,得到更精确的前景分割结果,其他图像也可以直接被推理实现前景分割。

图 1 展示了本文基于语义—表观特征融合的前景分割算法的完整框架。

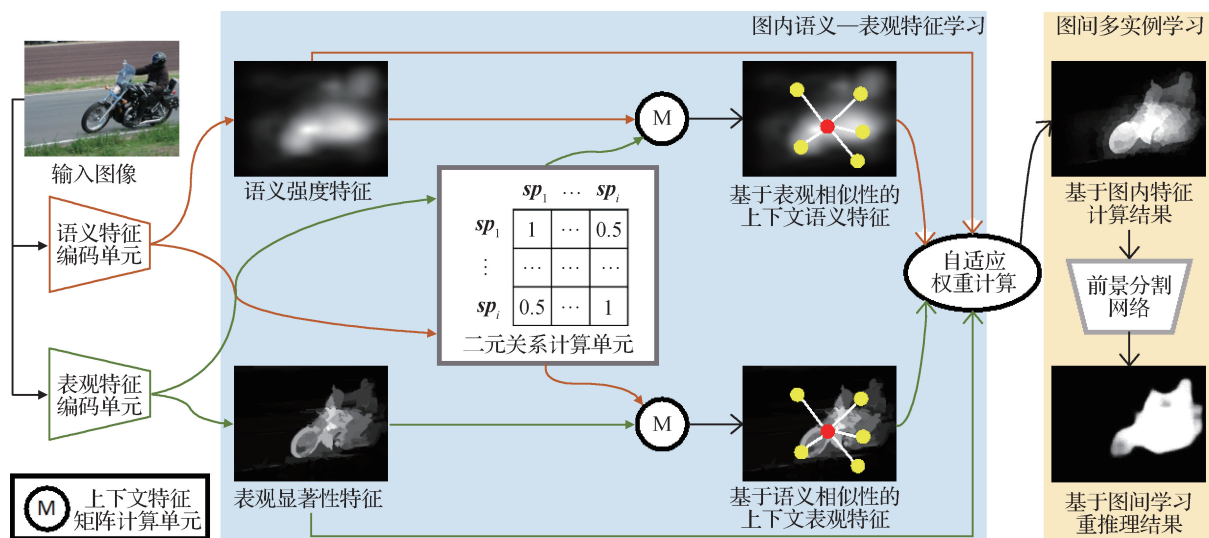


图 1 基于语义—表观特征融合的前景分割算法 (SAFF) 框架

Fig. 1 The framework of foreground segmentation method via semantic-apparent feature fusion (SAFF)

3 语义—表观特征编码

3.1 特征图生成及预处理

给定一幅图像,将其图像成像属性特征 (如颜

色、纹理、边缘和显著性等) 定义为表观特征,具有类别标签的特征定义为语义特征。为了编码图像的表观特征和语义特征,首先计算图像的语义特征响应图、显著性目标图和边缘图。

一方面,对于语义特征的生成,由于无监督条件

下没有针对目标数据集的额外标注信息,研究采用基于 ImageNet (Deng 等, 2009) 预训练的类别激活映射 (class activation mapping, CAM) (Zhou 等, 2016) 模型生成特征图。这里, GoogLeNet (Szegedy 等, 2015) 作为基础特征提取框架。对于一幅图像, 将编码得到 1 000 维的响应特征。需要说明, 研究并不像一般的弱监督语义分割任务一样, 需要训练集中的图像类别标签来训练 CAM 网络, 而是只将该模型作为语义特征提取器, 并没有使用 ImageNet 的类别信息, 也没有使用训练集中的其他标注信息。因此, 在任意数据集上进行训练时, 不再需要额外的标注信息, 故将任务定义为无监督前景分割任务。另一方面, 显著性响应图和边缘特征图用做表观特征。采用了判别性区域特征融合方法 (discriminative regional feature integration, DRFI) (Jiang 等, 2013) 进行显著性特征的计算, 而 EdgeBoxes (Zitnick 和 Dollár, 2014) 方法则用于提取边缘信息。需要说明的是, 任何一种高维语义热图、显著性、边缘的特征编码算法都可以用于语义和表观特征计算, 即本文方法可以广泛地将不同模型得到的特征作为初始信息, 进一步实现对图像中前景区域的更精准分割。

同时, 将图像进行了超像素分割, 并以超像素代表图像中最小的计算单元, 以同时提高局部特征的表达能力和保证计算的高效性。使用简单线性迭代聚类 (simple linear iterative clustering, SLIC) (Achanta 等, 2012) 方法, 通过设置初始的超像素数目, 在图像内均匀设置初始种子点, 并在种子点的邻域计算域内所有像素点的梯度值, 将种子点向梯度最小处移动。进一步地, 根据距离度量对每个种子邻域内的像素进行标签分类, 完成超像素分割过程。

利用特征编码结果, 定义了图像中每个超像素 sp_i 的语义、表观特征是该超像素包含的所有像素的特征值的均值。特别地, 语义特征图来自基于 CAM 的图像编码结果, vs_i 表示该超像素的 1 000 维归一化语义表达, 且各维度的值分布在 $[0, 1]$ 之间。表观特征中的显著性特征来自 DRFI 方法提供的显著性图, 研究中使用 $S_a(i)$ 表示其归一化的显著性特征。

3.2 一元区域特征编码

基于语义、表观特征, 首先针对每个超像素区域直接进行一元区域特征编码。在生成每种一元区域特征和二元上下文特征时, 将其值编码在 $[0, 1]$ 区

间内。具体地, 对于一个超像素 sp_i , 其一元区域语义特征定义为 vs_i 各个维度中的最大值 $S_s(i)$, 表观特征则直接以显著性特征 $S_a(i)$ 为代表。对于一幅包含 K 个超像素的样本而言, 2 个 K 维的特征向量 $[\dots, S_s(i), \dots]^T$ 和 $[\dots, S_a(i), \dots]^T$ 可以用来分别描述整幅图像的一元语义和表观特征。

3.3 二元上下文特征编码

利用语义特征表达, 对于一幅图像中的两两超像素对 sp_i 和 sp_j , 该超像素对语义相似度的表达为

$$M_s(i, j) = \sum_{d=1}^{1000} \min(vs_{i,d}, vs_{j,d}) \quad (1)$$

需要注意的是, vs_i 和 vs_j 是两个归一化的向量, 且各个维度的值均大于 0。在计算时, 本文对各维度取二者中的较小值, 而后采用加和的方式, 实际上是对两个向量的交叠面积进行了计算。因此, 得到的相似度计算结果是一个在 $[0, 1]$ 范围内的度量值, 值越大表明二者越相近。

另一方面, 参考了 Zheng 等人 (2015) 的研究, 超像素对之间的表观相似度定义为

$$M_a(i, j) = \exp(-w_e \cdot E(i, j)) \quad (2)$$

式中, 参考 AO-AIM (alternative optimization adaptive influence model) (Li 等, 2018c) 方法的定义, $E(i, j)$ 定义为超像素 sp_i 和 sp_j 之间的边缘重量。具体地, 利用 EdgeBoxes (Zitnick 和 Dollár, 2014) 方法提供的边缘图编码结果, $E(i, j)$ 定义为图像中两两超像素中心像素连线所经过的所有边缘图上像素值的和。 w_e 是一个正数, 在本文实验中, 参考了 Li 等人 (2018c, 2020) 的工作, 将其值设置为 3.5。容易得到, $M_s(i, j)$ 和 $M_a(i, j)$ 都是分布在 $[0, 1]$ 上的测度, 且值越大对应两者之间具有越高的特征相似度或越小的边缘图上距离。

利用基于语义相似度和表观相似度的计算方法, 图像中两两超像素对之间的相似度量可以表征为矩阵形式 M_s, M_a 。进一步地, 将其对角元置零, 并对矩阵逐行进行加和归一化处理, 得到修正后的矩阵 M'_s, M'_a 。对于每一个超像素 sp_i , 修正后的矩阵 M'_s 中的参数表示了图像中每个超像素与其他超像素的相似度量值。进一步地, 本文方法将该相似度量值作为权重, 以加权平均的方式, 定义了图中每个超像素的语义和表观上下文特征 $S_s^{ctx}(i), S_a^{ctx}(i)$ 。矩阵形式的全部超像素的上下文特征计算为

$$[\cdots, S_a^{ctx}(i), \cdots]^T = M'_s \cdot [\cdots, S_a(i), \cdots]^T \quad (3)$$

$$[\cdots, S_s^{ctx}(i), \cdots]^T = M'_a \cdot [\cdots, S_s(i), \cdots]^T \quad (4)$$

需要指出,在计算二元特征时,表观上下文特征是基于语义相似度矩阵和一元区域表观特征计算得到的;而语义上下文特征的生成则参考了表观相似度和一元语义特征。这种交叉特征推理的方式,一方面,更好地融合了两种特征并为前景区域推理提供了更丰富的表达;另一方面,避免了同一种特征同时被应用于一元表达和二元推理过程造成的误差累积。

4 自适应权重学习与推理

4.1 自适应权重学习

基于得到的一元、二元语义—表观编码特征,使用加权求和的方式,生成前景置信度图。对于每一个超像素 sp_i , 其置信分数 $Sco_{saff}(i)$ 定义为

$$Sco_{saff}(i) = Sco_v(i)^T \cdot w + bias \quad (5)$$

式中, w 为编码特征权重向量, $bias$ 为偏置系数, $Sco_v(i)$ 代表了该超像素编码得到的4元编码特征, 即

$$Sco_v(*) = [S_s, S_s^{ctx}, S_a, S_a^{ctx}]^T \quad (6)$$

考虑到对于不同的图像而言,一元、二元语义特征和表观特征对前景推理的影响不同,因此,本文并不把 w 和 $bias$ 设置成固定参数,而是采用自适应的方式实现参数的选取。首先,基于先验知识,假设当超像素的一元区域的表观、语义特征均具有高前景置信分数时应为前景,而均具有低置信分数时应判别为背景,并由此生成伪标注信息。具体地,给定一幅图像,对于每一个超像素 sp_i , 计算其一元表观特征和语义特征的几何平均值,并分别选取值小于 th_{bg} 和大于 th_{fg} 的超像素(bg 和 fg 分别代表了背景(background)与前景(foreground),将其置信分数设置为0和1作为伪标签(实验中, th_{bg} 和 th_{fg} 分别设置为0.2和0.6,该数值是在实验过程中,通过大量不同的参数设置实验后,选取的性能最佳的数值)。此时,将每一组结果代入式(5),可以得到以 w 和 $bias$ 为自变量的超定方程。此时,式(5)可以作为定义左右两端变量的目标函数,看做线性回归问题。这里,使用最小二乘方法,以得到对线性条件下最优参数的计算。在得到了权重 w 和偏置 $bias$ 后,利用

式(5)对图像中每一个超像素的前景置信度进行推理。这里限制分数的范围在 $[0, 1]$ 区间,对于区间外的结果,将其映射到相近的边界上。

在生成伪标签时,观察到一幅图像中选取出的背景样本往往远多于前景样本,而这会使得参数估计时对前背景置信分数计算产生一定偏差。因此,采用了一种样本均衡策略,即对较少的一类样本进行重采样,再将前背景平衡后的样本用于自适应参数估计的计算中。具体地,在对式(5)进行参数估计时,会选取 N_p 个对应于前景的超像素和 N_n 个对应于背景的超像素。对于其中较少的一类(例如 $N_p < N_n$),会在选取 N_p 个样本的基础上,进一步以等概率的方式在 N_p 个样本中采样 $N_n - N_p$ 次,实现正负样本选取数量的均衡化。最终,将超像素映射回图像的对应位置,可以得到一幅对应值分布在 $[0, 1]$ 区间的前景置信度图。

4.2 基于分割网络的重推理优化

对于每一幅图像,本文模型可以获取到融合了图像内部语义和表观上下文先验的前景区域推理结果。进一步地,基于 N 幅包含了不同场景和实例的图像,利用全监督前景分割网络学习的框架,实现对不同图像、多种实例中前景的共性特征表达学习,实现通过多实例学习过程提取前景共性特征。首先,对每幅图像对应的前景分割置信度图进行二值化操作,以0.5为阈值划分前景和背景,生成 N 张伪标签。利用伪标签,基于全监督学习的方式实现对分割网络的训练,本文采用以 VGG-16 (Visual Geometry Group 16-layer net) (Simonyan 和 Zisserman, 2014) 为基础网络的 DeepLab-V2 (Chen 等, 2018) 模型。事实上,该模块可以被任意语义分割或前景分割模型代替。利用训练完成的模型,每一幅图像可以被重新推理,并生成新的前景分割置信度图。同时,训练得到的网络模型也可以直接对其他图像进行前景分割的推理。通过融合基于图内语义—表观信息的上下文学习,和基于图间前景共性特征的多实例学习,可以获得更好的前景分割结果。

5 实验

5.1 数据集及训练和评估方法设置

基于 PASCAL VOC 2012 图像分割数据集

(Everingham 等, 2010) 进行了性能评估。使用增广数据 (Hariharan 等, 2011), 使训练集中共包含 10 582 幅图像样本, 验证集包含 1 449 幅图像样本。这一数据集包括了 20 类前景物体和 1 类背景, 评估时将所有属于物体类别的像素设置为前景。

进行网络训练时, 使用了以 VGG-16 为骨架的 DeepLab-V2 模型, 以语义分割标准的逐像素平均交叉熵损失函数作为优化对象。损失函数表达为

$$Loss = - \sum_{c=1}^M y_c \cdot \log(p_c) \quad (7)$$

式中, $Loss$ 为损失函数, M 是类别总数, y_c 只有 0 和 1 两种取值, 当对应像素类别为 c 设为 1, 否则为 0。 p_c 为网络预测输出的结果。

在训练时, 采用了 ImageNet 预训练后的特征提取网络参数, 并进行微调。基础学习率设置为 0.001, 学习率衰减采用了 Poly 模式, 采用 Momentum 作为优化器, 训练时设置的批处理大小为 16, 最大迭代次数为 20 000 次。训练好的网络参数将被固定, 并用于重推理训练集, 以及对其他任意测试图像实现前背景分割结果的输出。图 2 展示了训练过程的损失函数优化曲线。

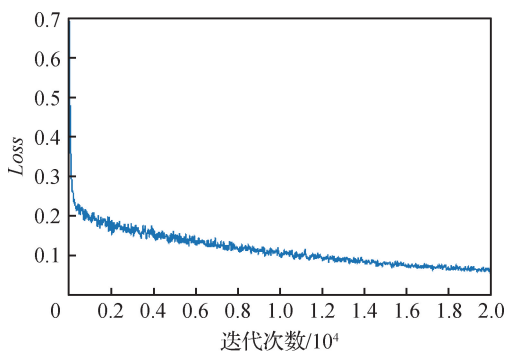


图2 训练过程优化曲线

Fig. 2 Optimization curve during training process

本文基于准确率 (precision)、召回率 (recall) 和 F-测度指标 (F-measure) 对前景分割的性能进行了评估。具体地, 对于每一幅图像和对应的前景置信度图, 本文将归一化的热图映射到从 0 ~ 255 的区间, 并分别计算以 0, 1, ..., 255 作为划分阈值时得到的准确率和召回率, 进一步地, 计算 F-测度指标

$$F_{\text{measure}} = \frac{(1 + \beta^2) \cdot P \times R}{\beta^2 \cdot P + R} \quad (8)$$

式中, P 和 R 分别表示准确率和召回率。参考典型的显著性区域检测、前景分割方法 (Li 等, 2018c; Wang 等, 2018b), 本文将 β^2 设置为 0.3。

5.2 实验性能及分析

针对 PASCAL VOC 2012 的训练集, 通过准确率—召回率曲线和 F-测度指标的对比来展示方法的性能。同时, 为了比较本文方法的性能并证明各个步骤的有效性, 对基线方法和本文方法中各个模块分步优化的结果都进行了评估。具体地, 对比了基于 DRFI (Jiang 等, 2013) 的显著性物体检测、基于 CAM (Zhou 等, 2016) 的语义热图生成的方法。在此基础上, 基于这些方法生成的伪标签, 再使用分割网络重推理优化的结果也被用于对比。同时进行了对不同模块的消融实验比较, 样本均衡处理、重推理优化模块被使用或舍弃时的算法性能均被评估。

需要指出, 本文针对的是无监督条件的前景分割问题, 因此并没有采用有专门任务标签训练条件下生成的语义和表观特征的方法作为基线模块。事实上, 任何一种语义特征和表观特征编码方法, 都可以用于本文提出的学习结构, 实现对前景分割任务性能的优化。

表 1 和表 2 分别展示了不同方法在 PASCAL VOC 2012 训练集和验证集上的 F-测度指标比较结果。除了基线方法外, 基于极弱监督的典型前景分割方法 (weaker supervised segmentation, WRSS) (Li 等, 2020) 也被陈列, 可以看到, 本文方法得到的结果与该典型方法相比, 仍达到了可比较的性能。在验证集上的结果是由以不同方法为伪标签训练的分割网络直接推理得到的, 并未对验证集图像进行其他操作。因此, 基于本文方法训练得到的模型进行前景分割的时间开销仅包括所使用的端到端分割网络推理时间, 可以高效地应用于实际分割任务。对于本文中使用的以 VGG-16 为骨架的 Deeplab-v2 模型而言, 算法的时间开销只与全监督语义分割网络的模型参数量和计算复杂度相关。能够看到, 本文方法通过自适应调整语义特征和表观特征对前景置信分数计算的贡献, 实现了更高性能的前景分割。另外, 在验证集上的结果也证明了基于 SAFF 得到的前景分割结果更好地指导了分割网络对前景提取的能力和在新数据上的泛化能力。

表 1 PASCAL VOC 2012 训练集前景分割 F-测度比较

Table 1 F-measure of foreground segmentation on PASCAL VOC 2012 train set

方法	样本均衡	网络重推理	F-测度
DRFI			0.638 3
DRFI		✓	0.651 3
CAM			0.604 4
CAM		✓	0.611 6
WRS		✓	0.684 1
SAFF (本文)			0.663 5
SAFF (本文)	✓		0.670 3
SAFF (本文)		✓	0.674 5
SAFF (本文)	✓	✓	0.686 3

注:加粗字体表示最优结果。

表 2 PASCAL VOC 2012 验证集前景分割 F-测度比较

Table 2 F-measure of foreground segmentation on PASCAL VOC 2012 validation set

方法	样本均衡	F-测度
DRFI		0.638 8
CAM		0.582 8
WRS		0.676 1
SAFF (本文)		0.659 2
SAFF (本文)	✓	0.672 5

注:加粗字体表示最优结果。

图 3 展示了不同方法的准确率—召回率曲线对比。除了基线以外,研究同样比较了在样本均衡、分割网络重推理策略使用与否条件下的性能结果。观察到,基于训练集、验证集的准确率—召回率曲线的评估结果同样证明了本文方法各个模块性能的有效性。图 4 和图 5 展示了部分在训练集和验证集上的可视化结果,可以看出,本文方法得到的前景分割同时收获了更好的语义区域定位和完整前景区域挖掘的性能。特别地,图 4 第 4 行的结果展示了通过图间多实例学习,初始的错误前景分割种子仍可以修正,并最终得到高质量的前景区域提取结果。

6 结 论

本文提出了一种基于表观—语义特征融合的无监督前景分割方法。对于一幅图像,首先计算图中

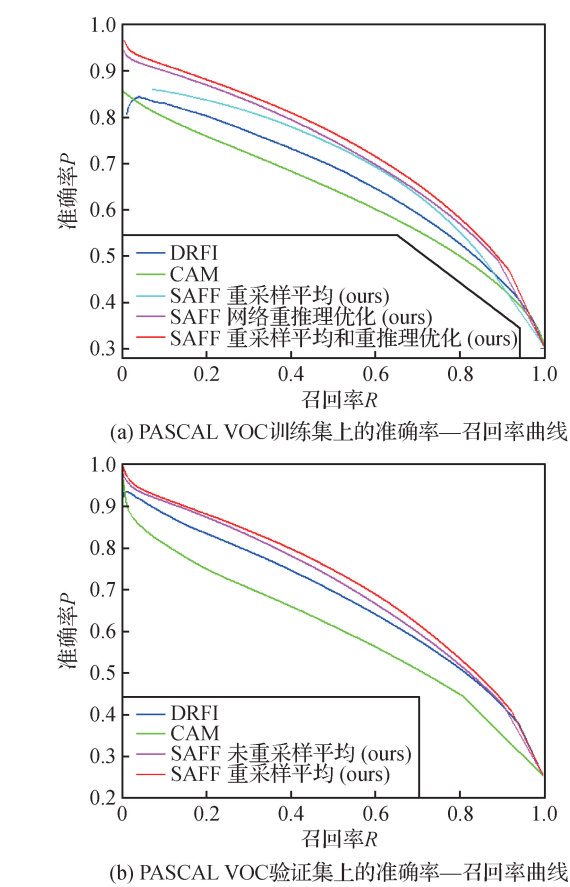


图 3 不同方法在数据集上的准确率—召回率曲线性能比较

Fig. 3 Comparison results of performance on precision-recall rate for different methods on dataset

((a)precision-recall rate on PASCAL VOC train set;

(b)precision-recall rate on PASCAL VOC validation set)

每个超像素的一元区域语义和表观特征。接着,通过交叉利用表观和语义特征的相似度量,实现了两种特征的融合,并计算得到了每个区域的二元图内上下文信息。进一步建立了自适应权重学习的策略,通过自动调节每一个特定图像实例中各个维度特征对前景估计的影响,得到一组最优的实现前景分割的权重参数,并实现了对图像前景置信度的推理。最后,使用分割网络模型以学习不同实例之间的前景共性特征。利用训练得到的网络模型,图像可以重推理得到更准确的前景分割结果。在 PASCAL VOC 2012 训练集和测试集上的实验分别证明了算法的有效性和泛化能力。同时,本文方法可将其其他前景分割方法作为基线,广泛用于对前景分割、弱监督语义分割等任务性能的提升。

在未来工作中,将考虑引入多种不同类型语义、表观特征融合,并采用交替迭代的方式,挖掘图像内

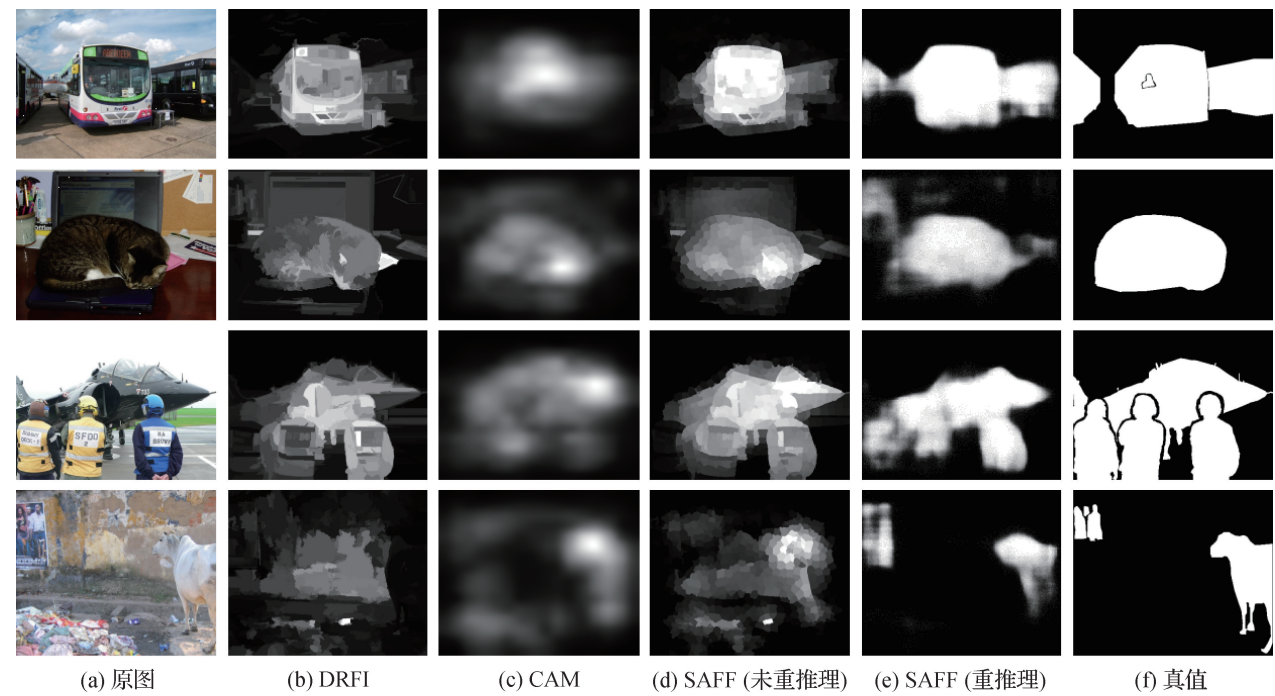


图 4 不同方法在训练集得到的可视化结果对比

Fig. 4 Visual results via different methods on train set

((a) original images; (b) DRFI; (c) CAM; (d) SAFF (no re-inferred); (e) SAFF (re-inferred); (f) truth)

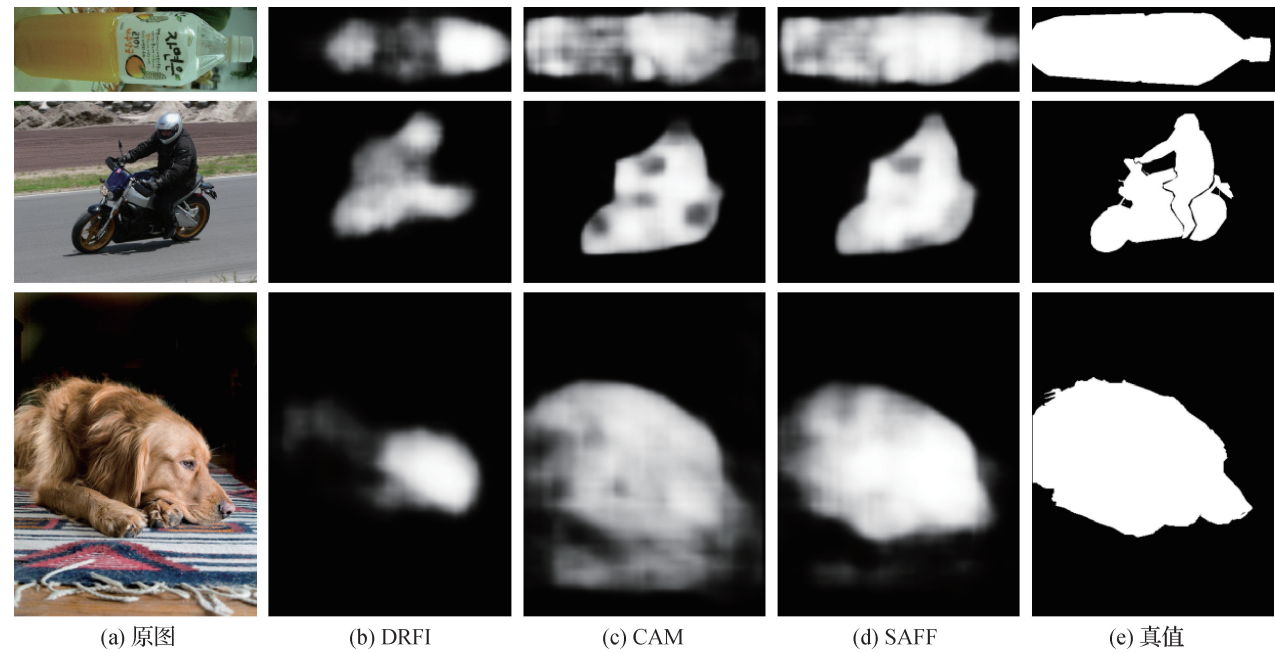


图 5 不同方法在验证集得到的可视化结果对比

Fig. 5 Visual results via different methods on validation set

((a) original images; (b) DRFI; (c) CAM; (d) SAFF; (e) truth)

部空间上下文信息和不同图像间的共性表达特征, 以实现更精准前景分割和语义分割。

参考文献 (References)

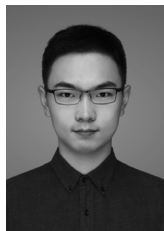
- Achanta R, Shaji A, Smith K, Lucchi A, Fua P and Süsstrunk S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2274-2282 [DOI: 10.1109/TPAMI.2012.120]
- Ahn J and Kwak S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA; IEEE: 4981-4990 [DOI: 10.1109/CVPR.2018.00523]
- Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]
- Chen L C, Papandreou G, Schroff F and Adam H. 2017. Rethinking atrous convolution for semantic image segmentation [EB/OL]. [2020-08-07] <https://arxiv.org/pdf/1706.05587.pdf>
- Chen X Z, Kundu K, Zhang Z Y, Ma H M, Fidler S and Urtasun R. 2016. Monocular 3D object detection for autonomous driving//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA; IEEE: 2147-2156 [DOI: 10.1109/CVPR.2016.236]
- Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//*Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, USA; IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]
- Everingham M, Van Gool L, Williams C K I, Winn J and Zisserman A. 2010. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2): 303-338 [DOI: 10.1007/s11263-009-0275-4]
- Hariharan B, Arbeláez P, Bourdev L, Maji S and Malik J. 2011. Semantic contours from inverse detectors//*Proceedings of 2011 International Conference on Computer Vision*. Barcelona, Spain; IEEE: 991-998 [DOI: 10.1109/ICCV.2011.6126343]
- Huang Z L, Wang X G, Wang J S, Liu W Y and Wang J D. 2018. Weakly-supervised semantic segmentation network with deep seeded region growing//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA; IEEE: 7014-7023 [DOI: 10.1109/CVPR.2018.00733]
- Jiang H Z, Wang J D, Yuan Z J, Wu Y, Zheng N N and Li S P. 2013. Salient object detection: a discriminative regional feature integration approach//*Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, USA; IEEE: 2083-2090 [DOI: 10.1109/CVPR.2013.271]
- Li X, Ma H M and Luo X. 2020. Weakly supervised semantic segmentation with only one image level annotation per category. *IEEE Transactions on Image Processing*, 29: 128-141 [DOI: 10.1109/TIP.2019.2930874]
- Li X, Ma H M and Wang X. 2018a. Feature proposal model on multidimensional data clustering and its application. *Pattern Recognition Letters*, 112: 41-48 [DOI: 10.1016/j.patrec.2018.05.025]
- Li X, Ma H M and Wang X. 2018b. Region proposal ranking via fusion feature for object detection//*Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*. Athens, Greece; IEEE: 1298-1302 [DOI: 10.1109/ICIP.2018.8451326]
- Li X, Ma H M, Wang X H and Zhang K. 2018c. Saliency detection via alternative optimization adaptive influence matrix model. *Pattern Recognition Letters*, 101: 29-36 [DOI: 10.1016/j.patrec.2017.11.006]
- Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA; IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]
- Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2020-08-07]. <https://arxiv.org/pdf/1409.1556.pdf>
- Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA; IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]
- Wang P Q, Chen P F, Yuan Y, Liu D, Huang Z H, Hou X D and Cottrell G. 2018a. Understanding convolution for semantic segmentation//*Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, USA; IEEE: 1451-1460 [DOI: 10.1109/WACV.2018.00163]
- Wang X, Ma H M, Chen X Z and You S D. 2018b. Edge preserving and multi-scale contextual neural network for salient object detection. *IEEE Transactions on Image Processing*, 27(1): 121-134 [DOI: 10.1109/TIP.2017.2756825]
- Wang X, You S D, Li X and Ma H M. 2018c. Weakly-supervised semantic segmentation by iteratively mining common object features//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA; IEEE: 1354-1362 [DOI: 10.1109/CVPR.2018.00147]
- Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z Z, Du D L, Huang C and Torr P H S. 2015. Conditional random fields as recurrent neural networks//*Proceedings of 2015 IEEE International*

Conference on Computer Vision. Santiago, Chile; IEEE: 1529-1537 [DOI: 10.1109/ICCV.2015.179]

Zhou B L, Khosla A, Lapedriza A, Oliva A and Torralba A. 2016. Learning deep features for discriminative localization//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE: 2921-2929 [DOI: 10.1109/CVPR.2016.319]

Zitnick C L and Dollár P. 2014. Edge boxes: locating object proposals from edges//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland; Springer: 391-405 [DOI: 10.1007/978-3-319-10602-1_26]

作者简介



李熹, 1994 年生, 男, 博士研究生, 主要研究方向为计算机视觉、弱监督学习、语义分割。
E-mail: lixi16@ mails. tsinghua. edu. cn



马惠敏, 通信作者, 女, 教授, 主要研究方向为计算机视觉、三维图像仿真、图像认知心理学。

E-mail: mhmpub@ ustb. edu. cn

马洪兵, 男, 研究员, 主要研究方向为遥感图像应用。

E-mail: hbma@ tsinghua. edu. cn

王弈冬, 男, 博士研究生, 主要研究方向为计算机视觉、目标检测、语义分割、图像认知心理学。

E-mail: wangyido@ mails. tsinghua. edu. cn