基于高阶 CRF 模型的图像语义分割*

毛 凌，解 梅

（电子科技大学 电子工程学院 图像处理与信息安全实验室，成都 611731）

摘 要： 图像语义分割方法大多基于点对条件随机场模型，不能定位到单个目标，并且难以利用全局形状特征，造成误识。针对这些问题，提出一种新的高阶条件随机场模型，将基于全局形状特征的目标检测结果和点对条件随机场模型统一在一个概率模型框架中，同时完成图像分割、目标检测与识别的任务。利用目标检测器和前背景分割算法获取图像中目标区域，在目标区域上定义新的高阶能量项。新的高阶条件随机场模型就是高阶能量项和点对条件随机场模型的加权混合模型，其最优解即为图像语义分割结果。在 MSRC-21 类数据库上进行的实验验证了该模型能够显著提升图像语义分割性能，并定位到单个目标。

关键词： 计算机视觉；图像语义分割；条件随机场模型；高阶能量项；基于可形变部件模型

中图分类号： TP391.4　　**文献标志码：** A　　**文章编号：** 1001-3695(2013)11-3514-04

doi:10.3969/j.issn.1001-3695.2013.11.081

Image semantic segmentation based on higher-order CRF model

MAO Ling, XIE Mei

(*Image Processing & Information Security Laboratory*, *School of Electric Engineering*, *University of Electronic Science & Technology of China*, *Chengdu* 611731, *China*)

Abstract: Current image semantic segmentation methods mostly use pairwise conditional random field (CRF) models, which can not distinguish instances of objects and tend to recognize wrongly for lack of global shape features. To solve of these problems, this paper proposed a new higher-order CRF model, which incorporated the pairwise CRF model and object detection based on global shape features into a unified probabilistic framework, and completed image segmentation, object detection and recognition tasks all at the once. It defined new higher-order energy terms on the object regions which were segmented out by the object detector and foreground-background segmentation algorithm. The proposed higher-order CRF model was the weighted combination of the higher-order energy terms and pairwise CRF model. The experiments conducted on the MSRC 21-class database show that the new higher-order CRF model can improve image semantic segmentation and locate instances of objects.

Key words: computer vision; image semantic segmentation; conditional random field models; higher-order energy term; deformable part models

0 引言

图像语义分割(image semantic segmentation)融合了传统的图像分割和目标识别两个任务，将图像分割成一组具有一定语义含义的块，并识别出每个分割块的类别，最终得到一幅具有逐像素语义标注的图像[1~4]，如图 1 所示。图 1 给出了图像语义分割研究中常用的典型数据库 MSRC-21 类数据库的示例图片及其语义标志图，其中第一行是原始图片，第二行是人工标志图，即真实数据(ground truth data)。不同颜色代表不同的目标类别，黑色表示空类；在训练和测试过程中忽略具有黑色标志的图片区域。目前的语义分割算法一般通过构建点对条件随机场模型(pairwise conditional random field models, pairwise CRFs)来融合图像分割和识别两个任务。点对 CRF 模型的一个优势是可以非常自然地加入图像局部纹理特征、全局上下文信息和平滑先验[1,3]，从而较好地完成图像分割和识别任务。

众所周知，局部纹理信息需要与全局形状特征互为补充，才能获得更好的识别效果，而点对 CRF 模型自身的特性导致很难引入全局形状特征，限制了识别率的提升。此外，基于点对 CRF 模型的方法无法区分单个目标，不能提供目标的数量信息。另一方面，目标检测方法大多利用了物体的全局形状特征，可以准确定位到单个物体[5]。然而，对于没有固定形状和空间区域的物体，如天空、道路，则无法检测，也无法得到目标的准确区域。

图1 图像语义分割

从上述考虑出发，本文提出一种新的高阶 CRF 模型，将基于全局形状特征的目标检测结果和点对 CRF 模型统一在一个概率模型框架中，同时完成图像分割、目标检测与识别的任务。

收稿日期： 2013-01-18；**修回日期：** 2013-03-07　　**基金项目：** 四川省科技支撑计划资助项目(2010GZ0153)

作者简介： 毛凌(1982-)，男，四川人，博士，主要研究方向为计算机视觉、机器学习(maoling5@163.com)；解梅(1955-)，女，教授，博士，主要研究方向为图像处理、模式识别.

本文的基本思路是:首先通过目标检测器准确定位输入图像中所有单个目标,然后从定位矩形框中提取出目标占据区域,再在这些区域上定义新的高阶能量项。新定义的高阶能量项与原始点对 CRF 模型的加权混合模型构成本文提出的高阶 CRF 模型,其最优解为语义分割结果。最后,将最优解的值带入到高阶能量项中,可以定位单个目标。

1　点对 CRF 模型

条件随机场是一种经典的判别型概率图模型,由 Lafferty 等人[6]于 2001 年正式提出,应用于文本分析。Kumar 等人[7]首先将其引入图像识别领域,利用 CRFs 构建了一个判别框架来识别复杂自然场景中的建筑物。文献[8]把从图像的局部和全局等不同尺度中获取的上下文信息融入该判别框架,将其推广应用到了含有七种目标类别的数据库上。Shotton 等人[1]提出一种基于纹元(texton)特征、空间先验、平滑先验与 JointBoost 分类器组合的点对 CRF 模型,在 MSRC-21 类数据库上取得了当时的最好结果。用于图像语义分割的经典点对 CRF 模型由式(1)给出,即已知输入图像时,语义分割结果 x 的后验概率表达式。点对 CRF 模型属于 MAP-MRF 模型。

$$P(x|y) \propto \exp(-(\sum_i E_i(x_i,y) + \alpha\sum_{ij} E_{ij}(x_i,x_j,y))) \quad (1)$$

如图 2 所示,应用于图像语义分割的典型点对 CRF 模型中[1],输入图像(图 3(a))宽为 W、高为 H,图像像素用观察变量 $y=\{y_1,y_2,\cdots,y_{W\times H}\}$ 来表示,像素的类别标志表示成隐随机变量 $x=\{x_1,x_2\cdots,x_{W\times H}\}$,从预先定义的类别集 $L=\{l_1,l_2,\cdots,l_n\}$ 中取值。在 MSRC-21 类数据库中,目标类别有自行车、行人、汽车、飞机等 21 类,即 $n=21$。图像语义分割的任务就是从类别集 L 中选择一个类别标志 l_i 分别分配给每个类别标志随机变量 x_i。这样的分配方式共有 $n^{W\times H}$ 种,需要从中选择出一种最符合人的感知的分配方式。在该模型中,满足最大后验概率的 x 就是最终的类别标志分配结果

$$x^* = \arg\max_x P(x|y) \quad (2)$$

式(1)中,根据含有隐随机变量的个数,E_i、E_{ij}分别称为一元势函数(unary potential)和点对势函数(pairwise potential)。E_i 定义在单个隐变量 x_i 和局部图像特征上,如图 2 中黄色(见电子版)连线所示,表示在观察到像素 y_i 所在局部区域时,x_i 取得某种类别标志的可能性,一般采用判别型分类器。E_{ij}定义在图像空间中相邻的成对隐变量 x_i 和 x_j 上,表示相邻像素取同种类别的可能性更大,因此也称为平滑势函数。这两种势函数的具体表达形式请参阅文献[1]。α 是 E_{ij}势函数的权重系数。

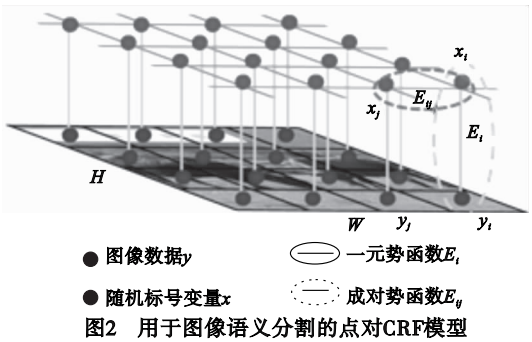

图2　用于图像语义分割的点对CRF模型

点对 CRF 模型一般只考虑局部区域的纹理特征,难以引入全局特征,尤其是全局形状特征,容易发生误识,并且平滑势函数易造成邻近目标区域发生粘连。比如采用该模型得到的语义分割结果图 3(b)中,两头牛中间的草地区域就因为过平滑而与牛所在区域发生粘连,加上缺乏形状特征而被误识成牛。该模型的另外一个局限性是不能定位到单个目标,比如根据粘连在一起识别为牛的区域,是无法判断出每头牛的准确位置和数目的。

图3　点对CRF模型的语义分割结果

2　基于可形变部件模型的目标检测

目标检测是计算机视觉领域的一个极具挑战性的问题。由于光照和视角变化,各类目标的外观变化非常大,基于固定形状特征和单一表观特征的检测器检测效果不佳,而非刚性形变、类间形状以及其他视觉特性差异进一步加剧了检测困难。Felzenszwalb 等人[5]提出一种基于多尺度可形变的部件模型(deformable part models,DPMs),考虑了非刚性形变因素,可以在一定程度上克服非刚性形变和视角变化造成的识别困难。

Felzenszwalb 等人提出的目标检测模型由一组可形变部件模板和一个全局模板组成。这些模板都由方向梯度直方图来表征,并在输入图像的不同尺度上计算得到,即在低分辨率上获得全局模板,高分辨率上获得部件模板。通过隐支持向量机(latent support vector machine,LSVM)对经过部件模型匹配后的响应特征进行学习以定位图像中的目标。该 DPMs 方法结合了全局形状特征和局部形状特征,具有良好的检测性能。图 4 中,(a)~(c)展示了文献[5]训练得到的牛的可形变模型;(d)是由该模型得到的实际检测结果,检测框左上角标注了识别类别和 LSVM 的响应值(信任度)。但是,DPMs 依然存在三个缺陷:a)只能检测具有区别性形状和固定空间区域的物体,如行人、车辆,无法检测不具备这两个条件的物体,如天空、草地、道路等;b)检测到的物体都使用矩形框定位,而矩形框内不可避免地会存在非物体区域,不能获得准确的物体区域;c)可能将同一个物体区域识别为不同类别的物体。

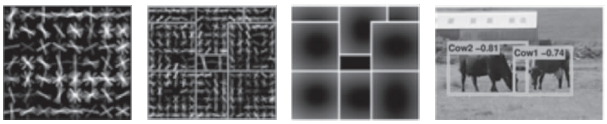

图4　可形变部件模型及其检测结果

考虑到上述两种方法各自的局限性,本文提出一种新的高阶 CRF 模型,将两种方法融合起来,结合了 DPMs 准确定位单个目标和点对 CRF 模型能够识别单个像素类别的优点,进而改善最终的图像语义分割结果,实现图像分割、目标检测与识别任务的统一。

3　本文方法

3.1　高阶 CRF 模型

为导出高阶 CRF 模型,先将第 1 章中给出的点对 CRF 模型后验概率形式转换成能量函数形式。由式(1)可知,求后验概率 $P(x|y)$的最大值,就是求 $\sum_i E_i(x_i,y) + \alpha\sum_{ij} E_{ij}(x_i,x_j,y)$的

最小值。据此,定义点对 CRF 模型的能量表达式为

$$E(x)=\sum_i E_i(x_i)+\alpha\sum_{ij}E_{ij}(x_i,x_j) \tag{3}$$

注意,这里忽略了观察变量 y。在能量表达式中,势函数 E_i 和 E_{ij} 也称为能量项(energy term)。式(2)等价于:

$$x^*=\arg\min_x E(x) \tag{4}$$

为了将 DPMs 的检测结果引入到 CRF 模型中,本文采用类似文献[9]中的方法,在 DPMs 检测得到的物体区域上定义新的高阶能量项:

$$E_{d_k}(x_{d_k})=-|x_{d_k}|\max(0,(1-R)\max(0,(C_{d_k}-C_t))) \tag{5}$$

$$R=\frac{N_{d_k}}{R_t|x_{d_k}|} \tag{6}$$

其中:x_{d_k}是组成单个物体区域的所有像素对应的随机标志变量的集合。因为其包含的随机变量数,即$|x_{d_k}|$远大于 2,所以称为高阶项。这个区域在图 5 中使用黄色透明色(见电子版)标记出来。第 3.2 节将介绍如何从矩形检测框中获取准确物体区域的方法。C_{d_k}表示 DPMs 对检测窗口区域的响应值(信任度);C_t 是阈值,通过调整该值可以控制最后的识别准确率;N_{d_k}是检测到的目标区域中标志类别与 DPMs 检测结果不同的像素数目;$R_t\in(0,1]$,本算法中取值范围在 0.1 ~0.3 时,效果较好。该参数对于物体遮挡或者物体区域提取不理想时,有较好的稳定作用。下角标 d_k 表示 DPMs 检测到的第 k 个目标,同时表示识别的类别号,$d_k\in L$。在第 1 章中的点对 CRF 模型中加入该高阶能量项,即得到新的高阶 CRF 模型,表达式为

$$E(x)=\sum_i E_i(x_i)+\alpha\sum_{ij}E_{ij}(x_i,x_j)+\beta\sum_k E_{d_k}(x_{dk}) \tag{7}$$

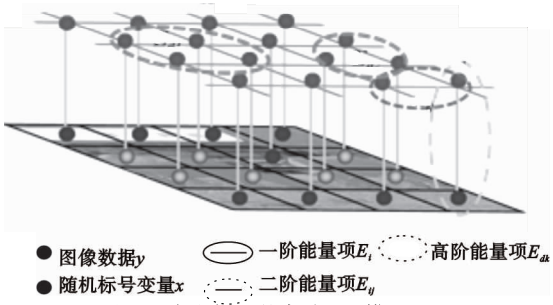

图5 本文提出的高阶CRF模型

模型中,高阶能量项的作用相当于惩罚项,即当 DPMs 检测到的物体类别与提取的物体区域内像素类别一致时,整体能量减小,且检测信任度越大,能量减小越多,反之能量增大。为了获得最小能量,在这些提取的物体区域上,原始点对 CRF 模型判断的类别应该与 DPMs 的判定尽可能保持一致,若不一致则取使能量更小的类别。因此,当基于形状特征的 DPMs 检测结果信任度很高时,就可能纠正基于纹理特征的点对 CRF 模型的错误识别。期望通过这种高阶能量项与非高阶能量项(E_i 和 E_{ij})的"竞争"(competition),可以获得更加准确的语义分割结果。图 6 展示了期望通过高阶能量项约束而获得的准确语义分割结果。

同 α 一样,β 是 E_{d_k}的权重系数,都是通过 Ladicky 等人[10]提出的贪婪算法在验证集上获得。最小化能量函数 $E(x)$的解 x 就是最后的语义分割结果。求解 x 的最优化方法有多种,本文采用 Kohli 等人[2]提出的基于图割(graph cuts)的最优化方法。要使用 Kohli 等人提出的最优化方法,高阶能量项必须满足一定条件。下面来证明式(5)满足该条件。

Kohli 等人指出,适合他们提出的最优化方法求解的势函数必须具备如下形式:

$$E_d(x_d)=\min(\min_{n\in L}(\gamma_n+R(\gamma_{\max}-\gamma_n)),\gamma_{\max}) \tag{8}$$

注意,原文献[2]中式(17)的变量符号已相应替换为本文使用的变量符号。为了简化推导过程,定义为

$$F=|x_{d_k}|\max(0,(C_{d_k}-C_t)) \tag{9}$$

现在,高阶能量项可以表示为

$$E_{d_k}(x_{d_k})=\min(0,-(1-R)F)=-F+\min(F,RF) \tag{10}$$

其中:$-F$ 是常数项,对最小化求解无影响。因此,只要证明 $\min(F,RF)$与式(8)形式上等价即可。显然,当设置满足式(11)中的条件时,这两者形式上等价。

$$\begin{aligned}\gamma_{\max}&=F\\ \gamma_n&=F,\ \forall n\neq d_k\\ \gamma_{d_k}&=0\end{aligned} \tag{11}$$

将最优解 x^* 代入$(1-R)\max(0,(C_{d_k}-C_t))$中,若大于 0,则认为此处存在单个物体。定位框可以利用 DPMs 得到的原始矩形检测框,也可以利用分割识别得到物体区域的外接矩形。这样就能够定位到单个目标,完成了目标检测任务。

3.2 物体区域分割

利用 DPMs 模型检测到的物体区域是一个矩形区域,里面总会存在非物体像素。若直接在矩形区域上定义高阶能量项,会影响最后的分割识别精度,所以需要去除这些非物体像素。本文采用 GrabCut[11]算法从检测到的矩形区域中提取出真实的物体区域。GrabCut 方法是对图割方法的改进,仅需要提供简单的手工标注信息(如矩形框),根据标注的信息对前景和背景建立高斯混合模型,最后利用图割方法求最优解。

图 7 是由 GrabCut 方法提取的物体区域。初始前背景信息由 DPMs 检测得到的矩形框提供(图 4(d))。可以看到,物体区域十分准确地分割出来。这部分区域是潜在的物体牛所在区域,在图 5 中使用黄色透明色(见电子版)覆盖,分别定义了两个高阶能量项 E_{d_1} 和 E_{d_2},为高阶 CRF 模型提供了区域类别一致性约束,与非高阶能量项"竞争"。但是,GrabCut 并不能准确地提取出每个检测框中的物体区域,有时可能只分割出物体的局部,尤其是当 DPMs 发生误检测时。所以,本文考虑当 GrabCut 获取的区域面积小于检测框面积的 50% 时,就忽略该检测框。

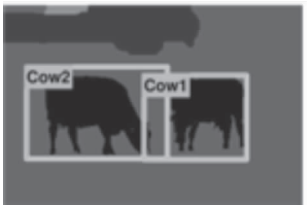

图6 期望的语义分割结果示意图

图7 GrabCut的分割结果

4 实验结果及分析

实验采用 MSRC-21 类数据库作为测试数据,并选择 Shotton[1]的方法作为比较对象。

MSRC-21 类数据库由微软剑桥研究中心发布,是目前最为复杂且手工标注较为完善的数据库之一。原始数据库由 591 幅图片组成,含有 23 类物体,其中马和山脉的图片非常少,一般都不考虑这两类。标注图中不同颜色代表不同的物体类别,黑色表示空类,在训练、测试中都被忽略。实验中,将该数据库图片随机分成训练集、验证集和测试集,各集占总图片

数的比例分别为 45%、10%、和 45%。

为保证比较的公平性,实验中高阶 CRF 模型的非高阶能量项 E_i 和 E_{ij} 的函数形式及函数参数的训练方法都与文献[1]保持一致,函数表达式和参数训练方法详见文献[1]。本文采用 Ladicky 等人[10]提出的贪婪算法在验证集上计算得到权重系数 α 和 β。由于文献[5]预先训练好的物体模板中只有 11 类物体与 MSRC-21 类数据库中物体种类重合,所以实验中只采用这 11 种检测模板来检测目标,实验结果对比分析也仅针对这 11 类物体进行。这 11 类物体参见图 8。

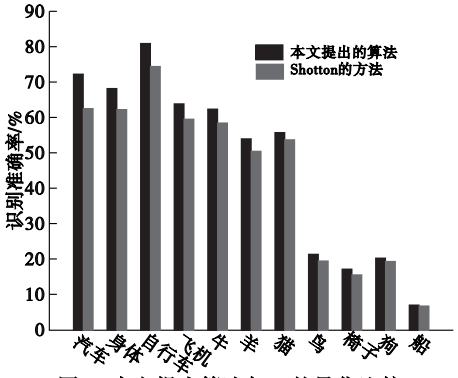

图8　本文提出算法与[1]的量化比较

下面,从定性和定量两个方面对这两种方法的实验结果进行比较分析。

a)从定性角度来验证本文提出的高阶 CRF 模型可以纠正点对 CRF 模型的错误识别。图 8 是实验结果对比图,其中,第一列为原始图像,第二列为 DPMs 检测结果,第三列为文献[1]的结果,第四列为本文提出算法的结果。可以看到,本文提出的高阶 CRF 模型由于结合了全局形状特征,可以纠正原始点对 CRF 模型的错误识别,并找到更加完整的物体区域。例如第一行中,两头牛中间原来被误识为牛的区域已正确识别为草地,不再粘连;与此类似,第二行牛的区域被准确识别出来,没有出现 Shotton 方法中误识粘连现象;第三行中,本文方法正确分割识别出了猫的完整区域;第四行和第五行中,点对 CRF 模型误识的人体区域也都在高阶 CRF 模型中得到了纠正。

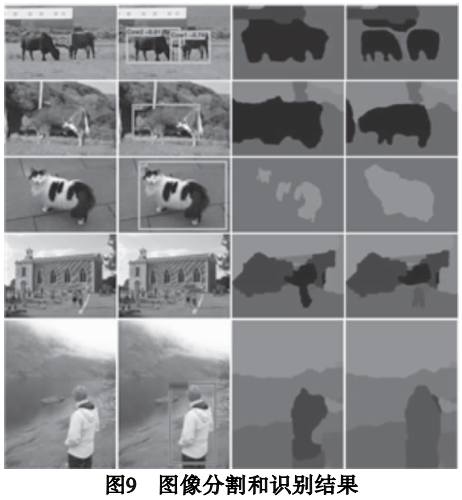

图9　图像分割和识别结果

需要指出的是,本文的模型对平滑能量项的过平滑作用解决得不理想,尽管 GrabCut 方法可以准确地提取出牛的边缘轮廓,但是最后分割得到的物体边缘部分不精确(请对比图 6、7 和 9 第一行结果)。此外,高阶能量项主要在 DPMs 检测框中发生作用,对检测框外区域的作用较小,这也是图 9 第一行结果中靠近建筑物附近的草地依然被误识为牛的原因。

b)为作量化对比,本文采用文献[1]提出的识别准确率(accuracy)作为评价标准,即正确识别样本与总样本之商。相比较 Shotton 的方法,本文方法对于图 9 中的 11 类物体都取得了更高的准确率,性能更为优越。其中,汽车、身体、自行车三类物体的准确率提高最为显著,分别提高了 10%、6.1% 和 6.3%;其余几类物体的识别准确率也有 1% ~4% 的提高。只有船这类物体提高不显著,低于 1%,这是由于 DPMs 模型对船的检测准确率偏低所致。

5　结束语

本文提出一种新的高阶 CRF 模型,将经典点对 CRF 模型和基于可形变部件模型的目标检测器统一在一个概率模型框架下,同时完成了图像分割、目标检测和识别的任务。在 MSRC-21 类数据库上的实验表明,本文提出的算法由于融合了点对 CRF 模型和基于可形变部件模型的目标检测器,结合了纹理特征和形状特征,与仅利用局部纹理信息的点对 CRF 模型相比,有效地提升了分割识别准确率,并且提供了物体的数量信息。

参考文献:

[1] SHOTTON J, WINN J, ROTHER C, *et al*. Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context[J]. International Journal of Computer Vision, 2009, 81(1): 2-23.

[2] KOHLI P, LADICK L, TORR P H S. Robust higher order potentials for enforcing label consistency[J]. International Journal of Computer Vision, 2009, 82(3): 302-324.

[3] GOULD S. Multiclass pixel labeling with non-local matching constraints[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washinton DC: IEEE Computer Society, 2012: 2783-2790.

[4] CARREIRA J, CASEIRO R, BATISTA J, *et al*. Semantic segmentation with second-order pooling[C]//Proc of the 12th European Conference on Computer Vision. Berlin: Springer-Verlag, 2012: 430-443.

[5] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, *et al*. Object detection with discriminatively trained part-based models[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[6] LAFFERTY J, McCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001: 282-289.

[7] KUMAR S, HEBERT M. Discriminative random fields: a discriminative framework for contextual interaction in classification[C]//Proc of the 9th IEEE IntenternationaL Conerence on Computer Vision. [S. l.]: IEEE Press, 2003: 1150-1157.

[8] HE Xu-ming, ZEMEL R S, CARREIRA-PERPINAN M A. Multiscale conditional random fields for image labeling[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2004: 695-703.

[9] GOULD S, GAO Tian-shi, KOLLER D. Region-based segmentation and object detection[C]//Proc of the 23rd Annual Conference on Neural Information Processing Systems. 2009: 655-663.

[10] LADICKY L U, RUSSELL C, KOHLI P, *et al*. Associative hierarchical CRFs for object class image segmentation[C]//Proc of the 12th IEEE International Conference on Computer Vision. [S. l.]: IEEE Press, 2009: 739-746.

[11] ROTHER C, KOLMOGOROV V, BLAKE A. GrabCut: interactive foreground extraction using iterated graph cuts[C]//Proc of ACM SIGGRAPH. New York: ACM Press, 2004: 309-314.