基于 RGB-D 的肉牛图像全卷积网络语义分割优化

邓寒冰[1,2]，周云成[1,2※]，许童羽[1,2]，苗 腾[1,2,3]，徐 静[1,2]

（1. 沈阳农业大学信息与电气工程学院，沈阳 110866；2. 辽宁省农业信息化工程技术研究中心，沈阳 110866；
3. 北京农业信息技术研究中心，北京 100097）

摘 要：基于卷积神经网络的深度学习模型已越来越多的应用于检测肉牛行为。利用卷积操作实现肉牛图像的像素级分割有助于实现远距离、无接触、自动化的检测肉牛行为，为肉牛异常行为早期发现提供必要手段。为了提高复杂背景下肉牛图像语义分割精度，降低上采样过程中的语义分割误差，该文提出基于 RGB-D 的肉牛图像全卷积网络（fully convolutional networks, FCN）的语义分割优化方法，用深度密度值来量化深度图像中不同像素点是否属于相同类型的概率，并根据深度图像与彩色图像在内容上的互补关系，优化和提升 FCN 对肉牛图像的语义分割（像素密集预测）精度。通过试验验证，该方法与全卷积网络的最优分割结果相比，可以将统计像素准确率平均提高 2.5%，类别平均准确率平均提升 2.3%，平均区域重合度平均提升 3.4%，频率加权区域重合度平均提升 2.7%。试验证明，该方法可以提升全卷积网络模型在复杂背景下肉牛图像语义分割精度。

关键词：图像处理；模型；动物；语义分割；RGB-D；全卷积网络；多模态；肉牛图像
doi：10.11975/j.issn.1002-6819.2019.18.019
中图分类号：S823.92; TP391.41 **文献标志码**：A **文章编号**：1002-6819(2019)-18-0151-10

邓寒冰，周云成，许童羽，苗 腾，徐 静. 基于 RGB-D 的肉牛图像全卷积网络语义分割优化[J]. 农业工程学报，2019，35(18)：151－160. doi：10.11975/j.issn.1002-6819.2019.18.019 http://www.tcsae.org
Deng Hanbing, Zhou Yuncheng, Xu Tongyu, Miao Teng, Xu Jing. Optimization of cattle's image semantics segmentation with fully convolutional networks based on RGB-D[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2019, 35(18): 151－160. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2019.18.019 http://www.tcsae.org

0 引 言

随着图像传感设备成本的不断降低，目前在畜牧养殖过程中已经逐步实现了全时段监控，特别是对动物全生命周期的行为监控和行为分析已经成为畜牧养殖业的一个研究热点。人们在获取大量动物图像和视频信息的同时，更关心如何实现对这些图像信息的处理、分析、理解和应用[1]；如何将动态的目标对象从复杂环境背景中分割出来，这是进行动物行为分析的前提条件，同时也是实现远距离、无接触、自动化检测动物行为的关键。

计算机视觉领域中的传统分割方法是通过人工提取图像特征来实现像素的聚类和提取，当图像背景复杂时，特征提取将变得非常麻烦甚至难以实现[2]。而随着深层卷积神经网络技术的发展，一种“端到端”的概念被引入到计算机视觉中来。让计算机自动在每个特定类别对象中学习和寻找最具描述性、最突出的特征，让深层网络去发现各种类型图像中的潜在模式[3]。在大量标注数据的基础上，通过不断的训练来自动提高卷积神经网络的分类、分割、识别、检测等处理的精度，将人工成本从算法设计转移到数据获取，降低了技术应用难度[4]。

在农业领域，基于卷积神经网络的计算机视觉技术已经逐渐成为主流研究方向。例如植物关键器官识别[5-8]，虫害个体识别[9-11]，植物病害分级[12-13]，利用多层卷积操作可以在不同尺度自动抽取图像特征，最后通过特征抽象可以获得目标位置和目标类型；针对家禽、水产等动物的视频图像处理方面，利用深层卷积网络可以实现针对动物个体轮廓提取、特征标定、行为轨迹追踪等[14-18]。然而，由于卷积神经网络中浅层的卷积感知域较小，只能学习到一些局部区域的特征；而深层的卷积层具有较大的感知域，对物体的大小、位置和方向等敏感性更低，有助于实现分类，但是因为丢失了物体的一些细节，不能指出每个像素具体属于哪个物体，很难做到精确的分割，不能够准确的给出目标对象的清晰边界信息[19-22]。而为了实现精准的像素分类，通常是以卷积过程中卷积核中心位置像素为基准点，通过判断该点周围区域像素组成的图像类别来预测该基准点的目标类别。然而，当卷积核区域不能覆盖一个完整对象时，预测精度会明显下降，而增大卷积核区域会造成运算过程中存储量的增加和计算效率的降低。为此，Evan 等提出了全卷积网络（fully convolutional networks，FCN）用于图像分割[23]，该网络从抽象的特征中恢复出每个像素所属的类别，与传统用 CNN 进行图像分割的方法相比，该网络采用的是全卷积

收稿日期：2019-04-02 修订日期：2019-08-20
基金项目：国家自然科学基金资助项目（31601218，61673281，31601219）；中国博士后科学基金（2018M631812）；辽宁省自然基金面上项目（20180551102）
作者简介：邓寒冰，讲师，博士，主要从事农业领域的机器学习与模式识别研究工作。Email：denghanbing@syau.edu.cn
※通信作者：周云成，副教授，博士，主要农业领域机器学习与模式识别研究工作。Email：zhouyc2002@syau.edu.cn

连接的结构，卷积过程共享感知区域，因此可以避免重复计算并提高卷积操作效率。

然而对于肉牛图像分割问题，由于肉牛所处的养殖环境复杂，图像中环境信息的颜色和纹理等会对肉牛形体细节部位的分割产生影响。特别是 FCN 在上采样过程中使用反卷积操作，对于图像中细节信息不敏感，没有考虑像素间的类别关系，使分割结果缺乏空间规整性和空间一致性[24]，这样得到的分割效果会非常粗糙。为了提高全卷积网络语义分割的精度，改善肉牛图像细节部位的分割效果，本文提出了基于 RGB-D 肉牛图像全卷积网络语义分割优化方法，定义了深度密度概念，利用深度密度值来量化深度图像中不同像素点是否属于相同类型的概率，并根据深度图像与彩色图像在像素内容上的映射关系，优化全卷积网络对肉牛图像的语义分割结果，提升分割的精度。

1 材料与方法

1.1 试验材料和准备工作

试验数据采集自辽宁省沈阳市北部地区肉牛养殖中心，肉牛品种为西门塔尔肉牛。为了增加样本多样性，试验在 5 月、8 月和 10 月，分别于上午（8:00－10:00）、中午（11:00－13:00）和下午（14:00－16:00）在室内和室外获取肉牛图像信息。采集设备为 Kinect Sensor（2.0 版本），可以同步获取分辨率为 1 920 像素×1 080 像素的彩色图像（RGB）和分辨率为 512 像素×424 像素的深度图像（Depth）。由于 Kinect 设备通过设备本身发出的结构光来计算物体的距离信息，所以在室外采集的深度图像存在较大的噪声，因此在室外采集过程中，只使用获取到的 RGB 图像（用于分类网络训练）；而室内采集的肉牛图像，由于外部光线可控，因此深度信息比较准确，可用于分割优化使用。在数据获取过程中，Kinect 设备位置固定，与拍摄对象（肉牛）保持 0.5～4.5m 距离，被拍摄对象在该范围可以自主活动。具体环境布局如图 1 所示。

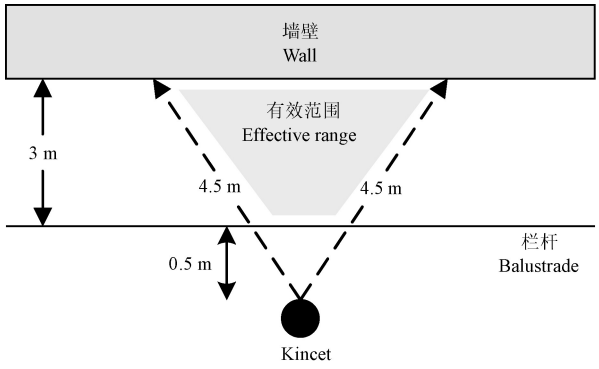


图 1 试验环境布局图

Fig.1 Layout of experiment environment

本试验选取的肉牛数量约为 70 头（室内 30 头，室外 40 头），从 Kinect 视频流等间隔（5 张/s）抽取 RGB 图像和深度图像，而且 RGB 图像和深度图像在时间轨迹上是同步的。将彩色图像通过人工加标注的方式形成 4 种用途的样本：用于分类网络训练，用于分类网络测试，用于分割网络训练和用于分割网络测试。在设定样本尺寸以及样本数量时，考虑到全卷积网络中不存在全连接层，因此可以实现对任意尺寸图片的处理。因此，本文利用可以将试验中用到的 RGB 图像和深度图像的尺寸统一到 512 像素×424 像素。为了增加样本多样性，分别于不同日期的上午、中午、下午 3 个时间段中各选取 2 000 张图像作为分类网络的训练样本（共 6 000 张），500 张图像作为分类网络的测试样本（共 1 500 张）；与此同时，在上述 3 个时间中，从室内采集的样本中选取 1 000 张分割网络的训练样本，200 张分割网络的测试样本。而深度图像是通过将 Kinect 获取的物体深度信息进行可视化表示后的效果图，即将可视范围内的深度值转换为灰度值，灰度归一化后范围是[0,1]，在后文中会利用深度图像计算每个像素点的深度密度，利用深度密度值来优化 FCN 语义分割结果。

本文后面章节将介绍如何设计试验和实现相关方法，具体包括 3 个主要过程，如图 2 所示。

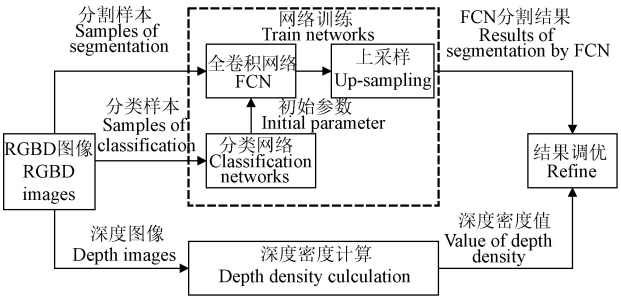


图 2 主要方法流程图

Fig.2 Flowchart of main method

1）利用分类样本集和分割样本集分别对基础分类网络和全卷积网络进行训练，其中基础分类网络的参数可以用于初始化全卷积网络，以加速训练过程中损失函数收敛；FCN 的输出特征图可以通过上采样得到初步的分割结果。

2）提出深度密度概念并给出深度密度计算方法，通过深度图像中每个像素点的深度密度，可以量化该像素点与周围空间其他像素点属于同一类别的概率。

3）利用深度密度值对分割结果中细节部位（例如边缘部位）进行调优，得到最终优化后的分割结果。

1.2 基础网络构建与训练

建立深层分类网络是解决逐像素预测问题和语义分割问题的基础，而 VGG 系列网络在 0～100 类左右的分类问题上，其分类精度与 Inception 系列、ResNet 系列等分类网络非常接近，而且 VGG 网络结构相对简单，没有 Inception 和 ResNet 网络结构中的用于优化训练的分支结构，因此更容易改造为全卷积网络，因此本文选择 VGG-19[25]作为分类网络的基本模型。VGG 系列网络在 ILSVRC2014（ImageNet[26] Large-Scale Visual Recognition Challenge）上首次提出，其网络结构参考了 AlexNet[27]。由于全卷积分割网络是在分类网络的基础上建立的，两

类网络在多个卷积层上是权值共享的，因此对分类网络进行预训练可以简化分割网络的训练过程，并且对分割精度有明显提升。此外，为了防止数据量不够而导致的过拟合问题，在训练分类网络的过程中加入了 ILSVRC2016 部分数据集，其中选择与试验场景相似的 150 类图片，形成了 151 类的数据集合。

在训练 VGG-19 方面，本文采用与文献[25]相同的训练方法。由于分类网络与全卷积分割网络在卷积层是共享权值的，因此在训练全卷积分割网络之前，对分类网络进行训练会提高分割网络的分割精度，同时缩短分割网络的训练时间。图 3a 给出了 VGG-19 训练过程中的损失函数的变化趋势图。由于本文使用的数据集规模要远小于 ImageNet[28]，因此在经历次 80 000 次 batch 迭代后，损失值已经在（0，0.05）之间，而平均分类精度可以达到 0.9 以上，已经基本达到了分类要求。同时基于同样的数据集和训练方法对 AlexNet 进行训练，训练结果如图 3b 所示。而相比 AlexNet 而言，虽然 VGG-19 层数更多，但是由于卷积核更小，因此网络收敛过程更加平稳，没有出现 AlexNet 在训练后期出现的 loss 值跳变的情况。所得到的 VGG-19 分类网络模型可以作为后文中全卷积网络的基础网络模型。

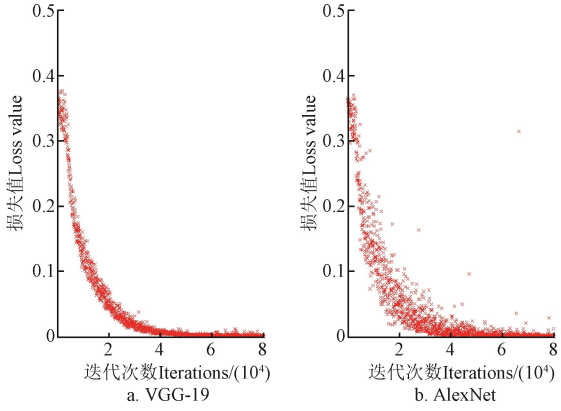


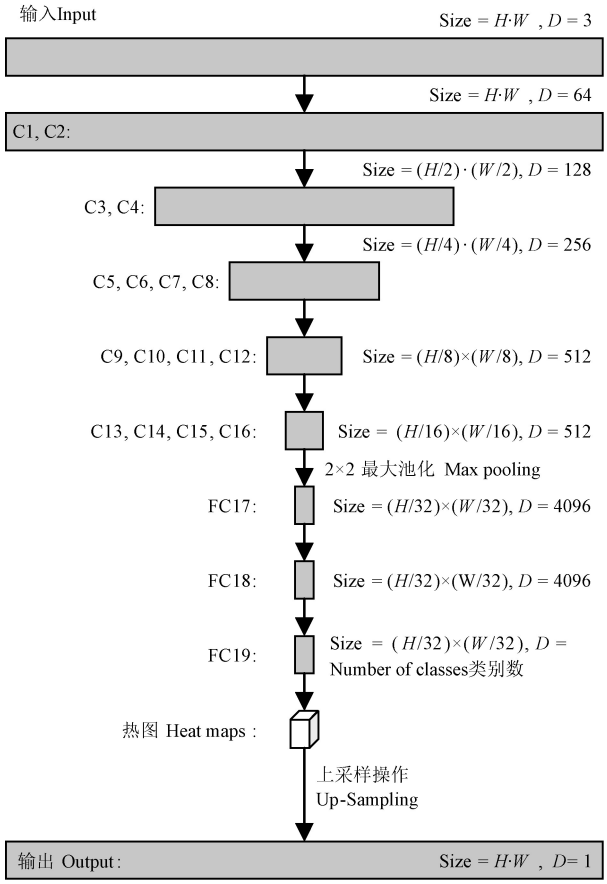
图 3　VGG-19 和 AlexNet 训练过程中的 loss 值

Fig.3　Training loss value of VGG-19 and AlexNet

1.3　全卷积网络构建

VGG-19 只能通过输出的特征向量来判定图像的具体类别，整个过程丢失大量的像素信息，无法实现像素级别的分类。全卷积网络（fully convolutional networks，FCN）以分类卷积神经网络为基础，将分类网络中的全连接层转换为卷积层，以保留输入图像的二维信息；对输出的特征图进行上采样操作，使其恢复到原始图像的尺寸，最后通过逐个像素分类获取每个像素的类别，从而实现对全图像的语义分割。本文中的全卷积网络的结构如图 4 所示，其中 H、W 分别表示初始图像的高和宽，D 表示通道数或维度。

FCN 是基于 VGG-19 建立的，每层池化操作可以使图片缩小一半，将 VGG-19 中的全连接层全部换成卷积层，卷积核的大小为 1×1，通道数保持不变，这样就可以保留特征图的二维空间属性，最终可以获得与类别数相等的热图（图 4 中 FC19）。热图的尺寸在经历过 5 次池化过程后，变成原图像大小的 1/32（如图 4 中 FC17、FC18 和 FC19）。为了实现端到端的语义分割，需要将热图进行 32 倍的上采样操作，以获取与原图像尺寸相等的语义分割结果。



注：C 表示卷积层，FC 表示卷积核为 1×1 的全卷积层；H 表示输入图像的高度值，W 表示输入图像的宽度值，D 表示输入图像和输出的特征图通道数。全卷积网络最后输出的特征图（热图）可以通过上采样操作获得与输入图像具有相同尺寸的语义分割结果。

Note: C denotes convolutional layer. FC denotes fully convolutional layer which has 1×1 convolutional kernel size. H denotes the value of image's height. W denotes the value of image's width. D denotes the number of channel for input image and output feature map. The final output (heat map) of fully convolutional networks can be transformed into semantic segmentation results as the same size as input image through up-sampling operation.

图 4　全卷积网络

Fig.4　Fully convolutional networks

1.4　基于全卷积网络的上采样操作

上采样（up-sample）是池化操作的逆过程，上采样后数据数量会增多。在计算机视觉领域，常用的上采样方法有 3 种：1）双线性插值[29]（bilinear）：这种方法特点是不需要进行学习，运行速度快，操作简单；2）反卷积[30]（deconvolution），利用转置卷积核的方法，对卷积核进行 180° 翻转；3）反池化[31]（depooling），在池化过程中记录坐标位置，然后根据之前坐标将元素填写进去，其他位置补 0。

与文献[23]中的上采样过程不同，为了提高上采样操作的精度，本文对于 2 倍尺寸的上采样操作选择双线性插值法，对于大于 2 倍尺寸的上采样操作选择反卷积法。对于双线性插值法，设原始特征图的尺寸为 $n\times n$，双线性插值法首先将原始特征图的尺寸变为$(2n+1)\times(2n+1)$，然

后利用 2×2 的卷积核对新特征图进行 valid 模式的卷积操作，最终获得尺寸为 2*n*×2*n* 的新特征图；而对于反卷积法，设原始尺寸为 *n*×*n*，利用 *m*×*m* 的卷积核对特征图进行 full 模式的卷积操作，最终可以获得尺寸为(*m*+*n*-1)×(*m*+*n*-1)的新特征图。

因为 VGG-19 中有 5 次池化操作，每经过一次池化操作，特征图的尺寸都变为原尺寸的 1/2。本文分别将每次池化后得到的特征图命名为 p1、p2、p3、p4 和 p5。如图 5 所示，输入图像的尺寸为 *H*·*W*，经过 5 次池化操作后 p5 的尺寸变为(*H*/32)·(*W*/32)。而 p1~p5 都可以作为本文上采样的输入特征图，参照输入图像的尺寸，分别恢复到对应特征图的 2 倍、4 倍、8 倍、16 倍和 32 倍。本文沿用文献[26]中的名称，称这些结果为 FCN-2s、FCN-4s、FCN-8s、FCN-16s 和 FCN-32s。（图 5 中只给出了 FCN-8s、FCN-16s 和 FCN-32s 的上采样过程）。为了解释计算过程，本文设输入图像的尺寸为 32×32 像素，VGG-19 网络中卷积操作不改变该阶段输入图像或特征图的大小，则 p1 的尺寸为 16×16 像素，p2 的尺寸为 8×8 像素，p3 的尺寸为 4×4 像素，p4 的尺寸为 2×2 像素，p5 的尺寸为 1×1 像素。FCN 最后的 3 个全卷积层的卷积操作（1×1 的卷积核）不会改变特征图的二维空间属性，因此输出的特征图尺寸仍与 p5 相等，为 1×1 像素，而且通道数与分类数（Number of classes）相等。

1）对于 FCN-32s，热图的大小为 1×1，FCN-32s 是由热图直接通过 32 倍的反卷积操作还原成 32×32 的尺寸。即用 *m*=32 的卷积核对 *n*=1 的特征图进行反卷积处理，输出的分割图为 32×32（*m*+*n*−1=32）。

2）对于 FCN-16s，对热图进行 1 次双线性插值操作，将热图的宽和高分别增大 2 倍，然后与 p4 相加，最后将相加的结果进行 16 倍的反卷积操作（*m*=31，*n*=2），可以获得与原图像相同尺寸的图像。

3）对于 FCN-8s，对热图进行 2 次双线性插值操作，使热图的宽和高分别增大 4 倍；然后对 p4 进行 1 次双线性插值操作，即将 p4 的宽和高分别增大 2 倍；最后将增大后的热图、p4 与 p3 相加，对相加的结果进行 8 倍的反卷积操作（*m*=29，*n*=4），可以获得与原图像相同尺寸的图像。

从结构上看，仍旧可以针对 p1 和 p2 的结果进行上采样处理，分别得到 FCN-2s 和 FCN-4s，但是根据文献[23]的结果显示在 8 倍上采样之后，优化效果已经不明显。因此，本文选择可以生成 FCN-8s 的全卷积网络作为语义分割的基础网络，但是上采样操作将热图中的分类像素点还原到原输入图像的尺寸，该过程存在较大的像素分类误差，即像素的错误分类以及像素丢失，而基于深度密度的图像分割优化方法可以用于优化该网络的语义分割结果。

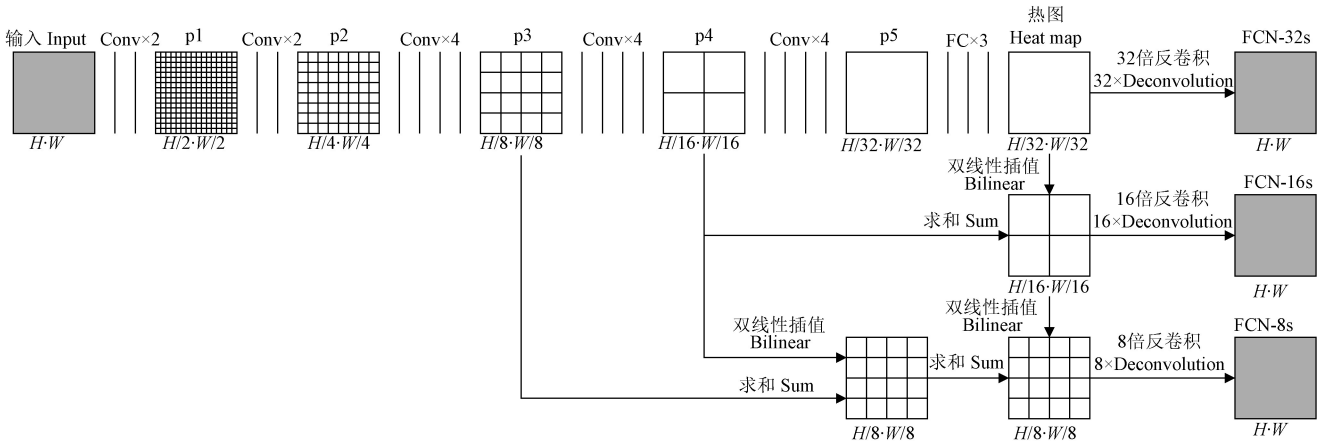


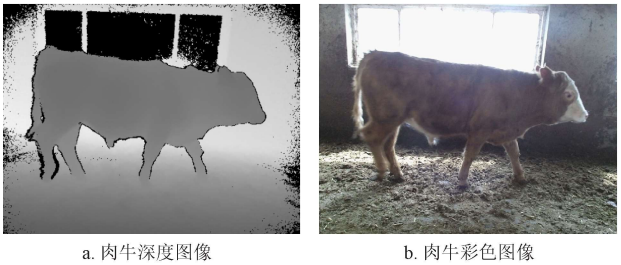
图 5 全卷积网络的上采样操作

Fig.5 Up-Sampling operation of fully convolutional networks

2 基于深度密度的图像分割优化

2.1 深度图像分析

深度图像中每个像素值表示空间中该点的位置与摄像头的空间距离，因此深度图像可以很好的描述复杂环境中肉牛的的轮廓信息（如图 6a 所示），而深度图像与 RGB 图像的像素之间存在内容上的映射关系（如图 6b）。在试验中，每张用于语义分割的 RGB 图像有与其对应的、具有相同尺寸的深度图像，而且通过 Kinect2.0 的软件处理，可以实现 RGB 图像与深度图像在内容上的近似映射。



a. 肉牛深度图像
a. Depth image of cattle

b. 肉牛彩色图像
b. RGB image of cattle

图 6 Kinect 获取的肉牛图像

Fig.6 Cattle images obtained by Kinect

从深度图像上可以看出，同一物体的细节信息可以通过连续变化的深度值表示出来，特别是对于同一目标来说，深度值一般是连续的，而相邻不同物体间的边界信息会出现深度值的跳变。通过统计可以发现，在同一张图片上，深度值相近的像素点在空间上有较大概率是是临近的，而且深度图像中属于同一物体并且在空间上连续的像素点，一般具有连续的灰度值区间。利用深度图像上的这一特点，本文提出了深度密度（depth density）的概念。

2.2　深度密度定义

设深度图像 I 的尺寸为 $r\times c$，其中 r 为图像 I 的行数，c 为图像 I 的列数；$dp(x,y)$为深度图像 I 上点(x,y)的深度值（由灰度表示）；$D(x,y)$表示图像 I 上点(x,y)对应的深度密度值，其表达式由公式（1）所示。

$$D(x,y)=f\left(dp(x,y),K^s_{x,y}\right) \tag{1}$$

式中 f 表示一个转换函数，其中参数 $K^s_{x,y}$ 表示以点(x,y)为中心，以 $s\times s$ 为范围的核区域。深度密度可以解释为：以深度图像 I 上(x,y)点为中心，计算该点与其 $K^s_{x,y}$ 区域内其他像素点的相似度。

为了计算相似度，本文首先给出几个参数定义：

1）$\mu^s_{x,y}$ 表示以(x,y)点为中心的 $K^s_{x,y}$ 区域内全部像素点的平均深度，即：

$$\mu^s_{x,y}=\frac{\sum_{i,j\in \boldsymbol{K}^s_{x,y}}\left|dp(i,j)\right|}{s^2} \tag{2}$$

式中 s 为深度密度计算过程中 K 区域边长。

2）$\sigma^s_{x,y}$ 表示以(x,y)点为中心的 $K^s_{x,y}$ 区域中，所有像素点与(x,y)点的深度方差均值，即

$$\sigma^s_{x,y}=\sqrt{\frac{\sum_{i,j\in \boldsymbol{K}^s_{x,y}}\left(dp(x,y)-dp(i,j)\right)^2}{s^2}} \tag{3}$$

3）$\sigma^m_{x,y}$ 表示以(x,y)点为中心的 $K^s_{x,y}$ 区域中，(x,y)点的深度值与该区域深度均值 $\mu^s_{x,y}$ 的方差，即

$$\sigma^m_{x,y}=\sqrt{\left(dp(x,y)-\mu^s_{x,y}\right)^2} \tag{4}$$

分别以 $dp(x,y)$ 和 $\mu^s_{x,y}$ 作为高斯分布函数 $X\sim N\left(\mu,\sigma^2\right)$的位置参数（$\mu$ 值），分别以 $\sigma^s_{x,y}$ 和 $\sigma^m_{x,y}$ 作为 $X\sim N\left(\mu,\sigma^2\right)$的尺度参数（$\sigma$ 值），这样可以获得 2 个正态分布函数，分别设为 g_s 和 g_m。

对于 $K^s_{x,y}$ 中每一个像素点的深度值，可以分别求出该点深度值在 $g_s\left(z,dp(x,y),\sigma^s_{x,y}\right)$ 和 $g_m\left(z,\mu^s_{x,y},\sigma^m_{x,y}\right)$ 中的概率密度值，这里分别用 $v^s_{i,j}$ 和 $v^m_{i,j}$ 表示，z 为函数 g_s 和 g_m 的深度自变量即

$$v^s_{i,j}=\int_{-\infty}^{dp(i,j)}g_s\left(z,dp(x,y),\sigma^s_{x,y}\right)dz \tag{5}$$

$$v^m_{i,j}=\int_{-\infty}^{dp(i,j)}g_m\left(z,\mu^s_{x,y},\sigma^m_{x,y}\right)dz \tag{6}$$

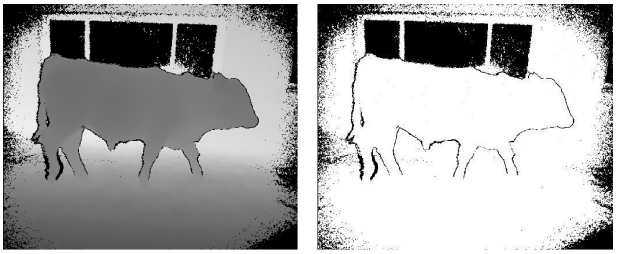
从 2 个函数的分布曲线可以分析，当 $v^s_{i,j}$ 和 $v^m_{i,j}$ 接近时，即表示 2 条曲线的分布重合度较高，证明 $K^s_{x,y}$ 区域中，中心点的深度值与区域的整体深度值分布情况比较接近，即区域中其他像素点与中心像素点有相近深度，属于深度密度较大的像素点；反之则表示该点的深度密度较小。本文设 $t(v_1,v_2)$来表示同一像素点在不同函数中深度分布的比值，即

$$t\left(v_1,v_2\right)=\begin{cases}v_1/v_2 & v_1<v_2\\ v_2/v_1 & v_1>v_2\end{cases} \tag{7}$$

这样，利用公式（8）就可以计算像素点(x,y)的深度密度值，即

$$f\left(dp(x,y),K^s_{x,y}\right)=\frac{\sum_{i,j\in \boldsymbol{K}^s_{x,y}}t\left(v^s_{i,j},v^m_{i,j}\right)}{s^2} \tag{8}$$

利用该公式计算得到的深度密度 $D(x,y)$的取值区间为(0,1]。其中，深度密度值越接近于 0 表示该点与该区域的整体深度值分布情况差异很大，则该点属于深度图中的边界像素或者噪声像素的概率较高；深度密度值越接近于 1 表示该点与该区域的整体深度值分布差异较小，则该像素点位于物体表面的几率较大。这就证明了如果一个像素点的深度密度接近于 1，则该点有很大概率与其周围 $s\times s$ 范围内的像素点属于同一物体。基于这一原理可以对全卷积的分割结果进行优化。图 7 给出了 s=7 时的深度密度图，其中图 7a 是肉牛的深度图像，图 7b 是深度图像通过计算深度密度计算后获取的深度密度图像。在深度密度图中，像素点的灰度值表示深度密度值，深度密度值越接近与 1（白色），表示该像素点与周围像素点深度值差别越小，而深度密度值越接近于 0（黑色），表示该像素点与周围像素点深度值差别越大，或该像素点在原深度图像中为无效小像素点。肉牛边缘处由于深度值变化明显、噪声多，因此边缘位置像素的深度密度值较低，而肉牛躯体部分由于深度值分布平滑，因此该位置深度密度值较高。



a. 深度图像　　b. 深度密度图像（s=7）
a. Depth image　　b. Depth density image (s=7)

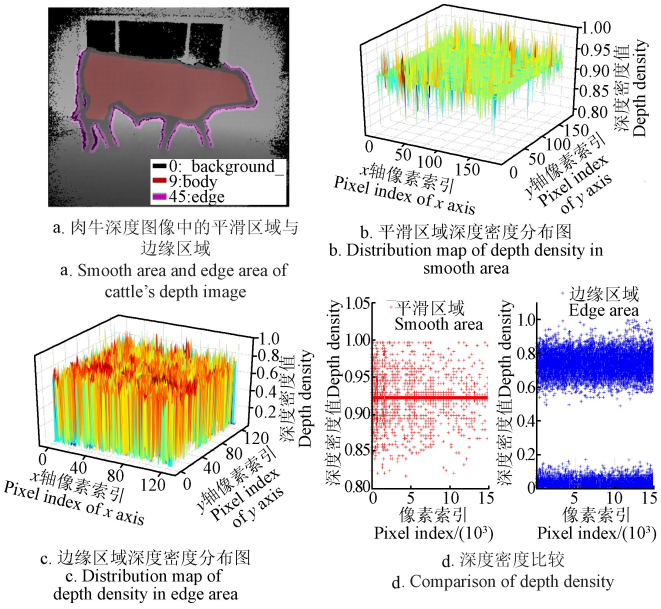
注：s 为深度密度计算过程中 K 区域边长。
Note: s is the edge length of K region in the calculation of depth density.

图 7　肉牛深度图像与深度密度图像
Fig.7　Depth image and depth density image

3　试验结果分析

试验利用深度密度对 FCN-8s 结果进行优化处理。根

据 2.2 节的公式（8），当 s 值较小时，对应的 $K^s_{x,y}$ 区域较小，深度密度算法的全局性较差，计算得到的深度密度只能反映局部区域的深度特征，在平滑区域会产生很多深度等高线，对小尺寸的噪声却很敏感，对尺寸较大的边界信息不是很敏感；而随着 s 值的增大，$K^s_{x,y}$ 区域也随着增大，算法的全局性有所改善，能够反映图像中更大区域的深度特征，深度等高线减少。但是 s 值较大时，算法对小尺寸的物体边界不敏感，这是由于核尺寸大于边缘宽度造成的，所以 s 值也不宜过大，而且尺度大会增加计算成本。本文根据深度密度图效果，选定 s=7 时来比较算法在深度图像中边缘区域与平滑区域的效果。



a. 肉牛深度图像中的平滑区域与边缘区域
a. Smooth area and edge area of cattle's depth image

b. 平滑区域深度密度分布图
b. Distribution map of depth density in smooth area

c. 边缘区域深度密度分布图
c. Distribution map of depth density in edge area

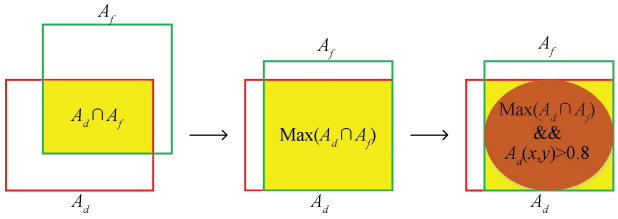
d. 深度密度比较
d. Comparison of depth density

图 8 平滑区域与边缘区域深度密度表示与对比

Fig.8 Depth density representation and comparison between smooth area and edge area

如图 8a 所示，在同一深度图中截取 2 类区域，其中红色区域（标注 9）表示深度图像中的深度平滑区域，粉色区域（标注 45）表示深度图像中深度边缘区域。通过像素映射找到 2 类区域对应的深度密度值，对这两个区域的深度密度进行分析。图 8 展示了 s=7 值条件下深度平滑区域的深度密度分布情况。其中图 8b 表示深度图像中平滑区域（图 8a 中红色区域）的深度密度值，该区域图的深度密度值普遍分布在[0.8, 1]区间，这表明该区域所在的像素点与其周围像素点的深度差非常小；而图 8c 表示深度图像中边缘区域（图 8a 中粉色区域），从图中可见，该区域深度密度值在[0, 0.8]区间反复震荡，这是由于深度图边缘区域深度值变化很大，同时 Kinect 采集的深度图像在物体边缘区域存在大量“黑色”噪点，因此边缘的深度密度值会更接近于 0 边缘区域也是产生噪声的主要区域，因此深度密度变化剧烈。图 8d 给出了平滑区域与边缘区域深度密度值的比较结果，其中分别在每个区域选取 15 000 个像素点进行比较，其中红色点表示平滑区域的深度密度值，蓝色点表示边缘区域深度密度值，从图中可以明显看到平滑区域像素点主要分布在[0.8,1]区间，而边缘区域虽然有些像素点的深度密度值也能达到 0.8，但那是由于在深度图像中截取边缘区域时附带的平滑区域像素点造成的。

综合考虑处理速度和处理效果，本文选取 s=7 作为深度密度的计算参数，将深度密度低于 0.8 的像素点过滤，在 FCN-8s 的基础上，可以得到更好的分割结果。图 9 为深度密度图与 FCN-8s 结果的融合过程，由于 A_d 区域为从深度图像中计算得到的，而 A_f 是 FCN 网络预测得到的，因此 A_d 区域有更高的分割可信度，因此首先在有限空间范围内（经过验证，调整区间为[-5,5]像素值）通过调整 A_f 位置来获取 $\mathbf{Max}\left(A_d \cap A_f\right)$；当获取 A_f 位置后，可以将 $A_d \cap A_f$ 区域内 $A_d(x,y) > 0.8$ 的像素点保留，则保留下来的区域即为 FCN-8s 优化后的结果，本文给该结果命名为 D-FCN-8s。



注：A_d 表示深度密度图像中的有边缘信息的对象区域，A_f 为 FCN-8s 结果中的分割区域。

Note: A_d denotes the object region with edges in depth density image. A_f denotes the segmentation region in FCN-8s.

图 9 深度密度图与 FCN-8s 融合过程

Fig.9 Fusion process of depth density image and FCN-8s

对于结果分析，本文选用 4 种通用的语义分割和场景解析的度量评价标准，用于评价像素精度和区域重合度，包括：统计像素准确率（pixel accuracy，pa）、类别平均准确率（mean accuracy，ma）、平均区域重合度（mean intersection over union，mIU）和频率加权区域重合度（frequency weight intersection over union，fwIU）。4 种评价标准的取值范围在 0 到 1 之间，值越接近于 1 表示分割精度越高。具体定义如下：

$$\mathrm{pa} = \sum_i n_{ii} \Big/ \sum_i t_i \tag{9}$$

$$\mathrm{ma} = \sum_i \frac{n_{ii}}{t_i} \Big/ n_{cl} \tag{10}$$

$$\mathrm{mIU} = \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \Big/ n_{cl} \tag{11}$$

$$\mathrm{fwIU} = \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \Big/ \sum_k t_k \tag{12}$$

式中 n_{ji} 表示属于 i 类别但被判别为 j 类的像素点个数，n_{cl} 表示像素总类别数，$t_i = \sum_j n_{ij}$ 表示 i 类像素点的总个数，k 表示类别数，n_{ii} 为正确识别的像素点个数。

为了避免对单一类别（肉牛）训练网络而造成的过拟合问题，本文将肉牛的训练数据与 NYUDv2 数据集[32]（40 个类别）混合使用，其中 NYUDv2 是用 Kinect 设备

采集并整理得到的一个公开 RGB-D 数据集合，该数据集中有 1 449 张 RGB-D 图像，同时包含 40 个类的语义分割标签。利用 NYUD2v 数据集，本文对 FCN 网络的 8 倍上采样语义分割结果（FCN-8s），以及 RGBD 图像语义分割结果（D-FCN-8s）在 4 种语义分割度量评价标准下进行精度比较。具体结果如表 1 所示。

表 1　在 NYUDv2 数据集上的语义分割比较

Table 1　Comparison of semantic segmentation on NYUDv2 data set

数据集 Data set	网络类型 Networks type	统计像素准确率 Pixel accuracy	类别平均准确率 Mean accuracy	平均区域重合度 Mean intersection over union	频率加权区域重合度 Frequency weight intersection over union
NYUDv2+1	FCN-8s	0.683	0.521	0.462	0.533
	D-FCN-8s	0.719	0.552	0.503	0.572
	精度差 D-value	0.036	0.031	0.041	0.039
NYUDv2-20+1	FCN-8s	0.862	0.578	0.496	0.773
	D-FCN-8s	0.883	0.604	0.531	0.795
	精度差 D-value	0.021	0.026	0.035	0.022
NYUDv2-10+1	FCN-8s	0.911	0.789	0.631	0.871
	D-FCN-8s	0.929	0.801	0.659	0.891
	精度差 D-value	0.018	0.012	0.028	0.020

注：FCN-8s 表示全卷积网络通过 8 倍上采样而获得的语义分割结果。D-FCN-8s 表示基于深度密度的全卷积网络通过 8 倍上采样而获得的语义分割结果。NYUDv2+1 表示在原 NYUDv2 数据集上添加 1 个新类别（肉牛）后形成的数据集（共 41 种类别）。NYUDv2-20+1 和 NYUDv2-10+1 同上。

Note: FCN-8s denotes the semantics segmentation result of fully convolutional networks by '8×' up-sampling. D-FCN-8s denotes the semantics segmentation results of fully convolutional networks based on depth density by '8×' up-sampling. NYUDv2+1 represents the data set (41 categories) formed by adding a new category (cattle) to the original NYUDv2 data set. NYUDv2-20+1 and NYUDv2-10+1 are the same with NYUDv2-40.

经过对比发现，当数据集类别减少时（41 类、21 类、11 类），FCN-8s 和 D-FCN-8s 在分割精度上都有一定的提升，这是因为全卷积网络的基础分类网络参数较多，而随着数据集类别的减少，网络训练过程出现了轻微的过拟合趋势。此外，使用 RGBD 图像进行语义分割时，通过判断深度图像中每个像素点的深度密度值是否操作特定阈值，可以区分该像素点是否处于肉牛边缘像素或肉牛躯体平滑区域，进而提高全卷积网络对 RGB 图像上采样语义分割的像素分类精度。参照表 1 中 D-FCN-8s 和 FCN-8s 对应的统计像素准确率（pa）、类平均准确率（ma）、平均区域重合度（mIU）和频率加权区域重合度（fwIU）的 4 组值，分别求得 D-FCN-8s 和 FCN-8s 在不同数据集（NYUDv2+1、NYUDv2-20+1 和 NYUDv2-10+1）下的精度差，最后可以求得平均精度差值（Average precision difference, APD），如表 2 所示，精度差值 D-FCN-8s 在统计像素准确率、类别平均准确率、平均区域重合度和频率加权区域重合度 4 种指标上比 FCN-8s 分别提高了 2.5%、2.3%、3.4%和 2.7%（表 2 中最后一列）。

为了验证该方法在 FCN 系列网络中的有效性，本文对原 FCN 的模型进行了改良，参照了文献[33]和文献[34]中的方案，在 FCN 结构后面加入了全连接条件随机场（conditional random fields, CRF）和马尔科夫随机场（Markov random fields, MRF），其中全连接条件随机场能够建立像素之间的全连接距离关系，而距离值与颜色和实际相对距离相关，这可以让该网络在语义分割过程中让图像尽量在边界处分割。而马尔科夫随机场对原 CRF 中的二元势函数进行了修改，加入了惩罚因子，能够更加充分的运用局部上下文信息产生分割结果。表 3 中给出了 4 种分割方案在 4 种通用的语义分割度量评价标准下的比较情况，其中 CRF-FCN-8s 是加入全连接条件随机场得到的语义分割结果，MRF-FCN-8s 是加入马尔科夫随机场得到的语义分割结果。结果表明，即时对原 FCN 网络进行改造，其各项指标也比深度密度对 FCN-8s 优化后的各项指标差，这是由于深度密度也采用了局部像素关联的方式来对具体像素点进行深度区域分类，而 CRF 和 MRF 虽然也是采用了距离关联方式，但是其关联关系的精度要低于深度图像中深度关联的精度，因此采用深度密度方法会得到更好的分割结果。这表明深度密度可以用于优化全卷积神经网络的语义分割结果，能够提升语义分割精度。图 10 分别给出 FCN-8s 以及为优化后的 D-FCN-8s 与真值的对比效果图，其中 FCN-8s 的分割细节部分明显不如 D-FCN-8s，而利用深度密度得到的分割结果非常接近与真值图。

表 2　FCN-8s 与 D-FCN-8s 在 3 类数据集上的平均精度差

Table 2　Average precision difference between FCN-8s and D-FCN-8s on three kinds of data set

数据集 Data set	统计像素准确率精度差 Precision difference of pa	类别平均准确率精度差 Precision difference of ma	平均区域重合度精度差 Precision difference of mIU	频率加权区域重合度精度差 Precision difference of fwIU
NYUDv2+1	0.036	0.031	0.041	0.039
NYUDv2-20+1	0.021	0.026	0.035	0.022
NYUDv2-10+1	0.018	0.012	0.028	0.020
APD	0.025	0.023	0.034	0.027

注：平均精度差（APD）的计算公式为，APD(average precision difference) = $((NYUDv2+1)_X+(NYUDv2-20+1)_X+(NYUDv2-10+1)_X)/3$，其中 $X \in \{pa, ma, mIU, fwIU\}$.

Note: Formula for calculating the average accuracy difference is as follows, APD (Average Precision Difference)=$((NYUDv2+1)_X+(NYUDv2-20+1)_X+(NYUDv2-10+1)_X)/3$, where $X \in \{pa, ma, mIU, fwIU\}$.

表 3　FCN-8s、CRF-FCN-8s、MRF-FCN-8s 和 D-FCN-8s 在 NYUDv2+1 数据集上的语义分割结果比较

Table 3　Comparison of semantic segmentation results of FCN-8s, CRF-FCN-8s, MRF-FCN-8s and D-FCN-8s on NYUDv2+1 data set

网络类型 Type of networks	统计像素准确率 Pixel accuracy	类别平均准确率 Mean accuracy	平均区域重合度 Mean IoU	频率加权区域重合度 Frequency weight IoU
FCN-8s	0.683	0.521	0.462	0.533
CRF-FCN-8s	0.686	0.533	0.477	0.542
MRF-FCN-8s	0.701	0.551	0.485	0.562
D-FCN-8s	0.719	0.552	0.503	0.572

注：CRF-FCN-8s 是以 FCN 为基础并加入全连接条件随机场后得到的分割结果，MRF-FCN-8s 是以 FCN 为基础并加入马尔科夫条件随机场后得到的分割结果。

Note: CRF-FCN-8s is s segmentation result based on FCN and adding Conditional Random Fields (CRF). MRF-FCN-8s is a segmentation result based on FCN and adding Markov Random Field (MRF).

图 10 D-FCN-8s、FCN-8s 与真值对比

Fig.10 Comparison with D-FCN-8s, FCN-8s and ground-truth

5 结 论

1）在对全卷积网络输出的特征图（热图）进行上采样过程中，交替使用了双线性插值方法和全尺寸反卷积方法，避免了直接采用全尺寸反卷积操作而造成的分割结果粗糙的问题。

2）利用深度图像，计算每个像素点的深度密度值，该值由深度核区域中其他像素深度值与中心像素点深度值关系决定，利用深度密度值可以量化该像素点与 $K^s_{x,y}$ 区域其他像素点属于同一像素类型的概率，通过试验分析，当深度密度值大于 0.8 时，属于同一类型的概率较大。

3）基于像素密度值，可以对 FCN-8s 中肉牛细节部分（例如边缘部位）进行优化，经过试验结果分析，在 3 类数据集上（NYUDv2+1，NYUDv2-20+1，NYUDv2-10+1）进行分割验证，与原始 FCN-8s 分割结果相比，D-FCN-8s 在统计像素准确率提高 2.5%，在类别平均准确率提升 2.3%，在平均区域重合度提升 3.4%，在频率加权区域重合度提升 2.7%。

4）本文在 FCN 的基础上，分别加入了全连接条件随机场和马尔科夫随机场，用于在对像素分类过程中增加像素局部上下文信息，提高 FCN 系列网络的分割精度，通过 NYUDv2+1 数据集验证发现 D-FCN-8s 结果仍优于这两种网络，因为深度密度是在深度图像中使用了局部深度全局信息，而深度图像的精度要高于全连接条件随机场和马尔科夫随机场中的距离值，因此分割效果更好。

因此，上述结论证明通过计算和使用 RGBD 图像中像素点的深度密度，可以优化全卷积网络在肉牛细节部位的分割效果，提高全卷积网络的语义分割精度。

[参 考 文 献]

[1] Zhu Nanyang, Liu Xu, Liu Ziqian, et al. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities[J]. IJABE. 2018, 1(4): 32－44.

[2] David Stutz, Alexander Hermans, Bastian Leibe. Superpixels: An evaluation of the state-of-the-art[J]. Computer Vision and Image Understanding. 2018, 166: 1－27.

[3] Bell S, Zitnick C L, Bala K, et al. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vega, 2016: 2874－2883.

[4] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2016, 313(5786): 504－507.

[5] 周云成，许童羽，郑伟，等. 基于深度卷积神经网络的番茄主要器官分类识别[J]. 农业工程学报，2017，33(15)：219－226.
Zhou Yuncheng, Xu Tongyu, Zheng Wei, et al. Classification and recognition approaches of tomato main organs based on DCNN[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(15): 219－226. (in Chinese with English abstract)

[6] 田有文，程怡，王小奇，等. 基于高光谱成像的苹果虫伤缺陷与果梗/花萼识别方法[J]. 农业工程学报，2015，31(4)：325－331.
Tian Youwen, Cheng Yi, Wang Xiaoqi, et al. Recognition method of insect damage and stem/calyx on apple based on hyperspectral imaging[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(4): 325－331. (in Chinese with English abstract)

[7] 赵源深，贡亮，周斌，等. 番茄采摘机器人非颜色编码化目标识别算法研究[J]. 农业机械学报，2016，47(7)：1－7.
Zhao Yuanshen, Gong Liang, Zhou Bin, et al. Object recognition algorithm of tomato harvesting robot using non-color coding approach[J]. Transactions of the Chinese Society for Agricultural Engineering, 2016, 47(7): 1－7. (in Chinese with English abstract)

[8] 贾伟宽，赵德安，刘晓样，等. 机器人采摘苹果果实的 K-means 和 GA-RBF-LMS 神经网络识别[J]. 农业工程学报，2015，31(18)：175－183.
Jia Weikuan, Zhao Dean, Liu Xiaoyang, et al. Apple recognition based on K-means and GA-RBF-LMS neural network applicated in harvesting robot[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(18): 175－183. (in Chinese with English abstract)

[9] 杨国国，鲍一丹，刘子毅，等. 基于图像显著性分析与卷

积神经网络的茶园害虫定位与识别[J]. 农业工程学报，2017，33(6)：156－162.
Yang Guoguo, Bao Yidan, Liu Ziyi, et al. Localization and recognition of pests in tea plantation based on image saliency analysis and convolutional neural network[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(6): 156－162. (in Chinese with English abstract)

[10] 谭文学，赵春江，吴华瑞，等. 基于弹性动量深度学习的果体病例图像识别[J]. 农业机械学报，2015，46(1)：20－25.
Tan Wenxue, Zhao Chunjiang, Wu Huarui, et al. A deep learning network for recognizing fruit pathologic images based on flexible momentum[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(1): 20－25. (in Chinese with English abstract)

[11] 王献锋，张善文，王震，等. 基于叶片图像和环境信息的黄瓜病害识别方法[J]. 农业工程学报，2014，30(14)：148－153.
Wang Xianfeng, Zhang Shanwen, Wang Zhen, et al. Recognition of cucumber diseases based on leaf image and environmental information[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2014, 30(14): 148－153. (in Chinese with English abstract)

[12] 王新忠，韩旭，毛罕平. 基于吊蔓绳的温室番茄主茎秆视觉识别[J]. 农业工程学报，2012，28(21)：135－141.
Wang Xinzhong, Han Xu, Mao Hanping. Vision-based detection of tomato main stem in greenhouse with red rope[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2012, 28(21): 135－241. (in Chinese with English abstract)

[13] 郭艾侠，熊俊涛，肖德琴，等. 融合 Harris 与 SIFT 算法的荔枝采摘点计算与立体匹配[J]. 农业机械学报，2015，46(12)：11－17.
Guo Aixia, Xiong Juntao, Xiao Deqin, et al. Computation of picking point of litchi and its binocular stereo matching based on combined algorithms of Harris and SIFT[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(12): 11－17. (in Chinese with English abstract)

[14] 赵凯旋，何东键. 基于卷积神经网络的奶牛个体身份识别方法[J]. 农业工程学报，2015，31(5)：181－187.
Zhao Kaixuan, He Dongjian. Recognition of individual dairy cattle based on convolutional neural networks[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(5): 181－187. (in Chinese with English abstract)

[15] 段延娥，李道亮，李振波，等. 基于计算机视觉的水产动物视觉特征测量研究综述[J]. 农业工程学报，2015，31(15)：1－11.
Duan Yan'e, Li Daoliang, Li Zhenbo, et al. Review on visual characteristic measurement research of aquatic animals based on computer vision[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(15): 1－11. (in Chinese with English abstract).

[16] 高云，郁厚安，雷明刚，等. 基于头尾定位的群猪运动轨迹追踪[J]. 农业工程学报，2017，33(2)：220－226.
Gao Yun, Yu Houan, Lei Minggang, et al. Trajectory tracking for group housed pigs based on locations of head/tail[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(2): 220－226. (in Chinese with English abstract)

[17] 邓寒冰，许童羽，周云成，等. 基于 DRGB 的运动中肉牛形体部位识别[J]. 农业工程学报，2018，34(5)：166－175.
Deng Hanbing, Xu Tongyu, Zhou Yuncheng, et al. Body shape parts recognition of moving cattle based on DRGB[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(5): 166－175. (in Chinese with English abstract)

[18] 杨阿庆，薛月菊，黄华盛，等. 基于全卷积网络的哺乳母猪图像分割[J]. 农业工程学报，2017，33(23)：219－225.
Yang Aqing, Xue Yueju, Huang Huasheng, et al. Lactating sow image segmentation based on fully convolutional networks[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(23): 219－225. (in Chinese with English abstract)

[19] 郭祥云，台海江. 深度学习在大田种植中的应用及展望[J]. 中国农业大学学报，2019，24(1)：119－129.
Guo Xiangyun, Tai Haijiang. Current situation and prospect of deep learning application in field planting[J]. Journal of China Agricultural University, 2019, 24(1): 119－129. (in Chinese with English abstract)

[20] 王丹丹，何东健. 基于 R-FCN 深度卷积神经网络的机器人疏果前苹果目标的识别[J]. 农业工程学报，2019，35(3)：156－163.
Wang Dandan, He Dongjian. Recognition of apple targets before fruits thinning by robot based on R-FCN deep convolution neural network[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2019, 35(3): 156－163. (in Chinese with English abstract)

[21] 刘立波，程晓龙，赖军臣. 基于改进全卷积网络的棉花冠层图像分割方法[J]. 农业工程学报，2018，34(12)：193－201.
Liu Libo, Cheng Xiaolong, Lai Junchen. Segmentation method for cotton canopy image based on improved fully convolutional network model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(12): 193－201. (in Chinese with English abstract)

[22] 段凌凤，熊雄，刘谦，等. 基于深度全卷积神经网络的大田稻穗分割[J]. 农业工程学报，2018，34(12)：202－209.
Duan Lingfeng, Xiong Xiong, Liu Qian, et al. Field rice panicle segmentation based on deep full convolutional neural network[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(12): 202－209. (in Chinese with English abstract)

[23] Evan Shelhamer, Jonathan Long, Trevor Darrell. Fully Convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640－651.

[24] Ronghang Hu, Piotr Dollar, Kaiming He, et al. Learning to segment every thing[C]//IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018, 4233－4241.

[25] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition[C]// International Conference on Learning Representations, San Diego, 2014: 1－14.

[26] Deng Jia, Dong Wei, Socher Richard, et al. ImageNet: A large-scale hierarchical image database[C]//IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2009: 248－255.

[27] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks[J]. International Conference on Neural Information Processing System, 2012, 60(2): 1097－1105.

[28] Jia Deng, Wei Dong, Richard Socher, et al. ImageNet: A large-scale hierarchical image database[C]// IEEE Conference on Computer Vision & Pattern Recognition, 2009: 248－255.

[29] Lin Tsungyu, Aruni RoyChowdhury, Subhransu Maji. Bilinear CNN models for fine-grained visual recognition[J].

IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1309－1322.
[30] Zheng Shou, Jonathan Chan, Alireza Zareian, et al. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1417－1426.
[31] Matthew D Zeiler, Rob Fergus. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision, Zurich, 2014: 818－833.
[32] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[C]//In ECCV, 2012.7
[33] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks[C]// 2015 IEEE International Conference on Computer Vision. 2015.
[34] Liu Ziwei, Li Xiaoxiao, Luo Ping, et al. Semantic image segmentation via deep parsing network[C]// IEEE International Conference on Computer Vision. 2015.

Optimization of cattle's image semantics segmentation with fully convolutional networks based on RGB-D

Deng Hanbing[1,2], Zhou Yuncheng[1,2※], Xu Tongyu[1,2], Miao Teng[1,2,3], Xu Jing[1,2]

(1. *College of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang* 110866, *China*;
2. *Liaoning Engineering Research Center for Information Technology in Agricultural, Shenyang* 110866, *China*;
3. *Beijing Research Center for Information Technology in Agricultural, Beijing* 100097, *China*)

Abstract: With the decreasing cost of image sensor equipment, full-time monitoring has been gradually realized in the process of cattle breeding. Especially, in the whole life of cattle, the monitoring and analysis for cattle's behavior have become a research hotspot in the field of breeding. Acquiring a large amount of cattle image and video information, people are more concerned about how to process, analyze, understand and apply these data. How to segment dynamic objects from complex environment background is the precondition of cattle behavior analysis, and it is also the key of realizing long-distance, contactless and automatic detection for cattle behavior. The traditional machine vision image segmentation method is used to realize the clustering and extraction of pixels by artificially extracting image features. However, when the image background is complex, feature extraction will become very troublesome and even difficult to achieve. Deep Convolutional Neural Networks (DCNN) provides another solution, which enables computers to automatically learn and find the most descriptive and prominent features in each specific category of objects, and allows deep networks to discover potential patterns in various types of images. On the basis of massive labeled data, the accuracy of classification, segmentation, recognition and detection with convolutional neural network can be improved automatically through continuous training, and the labor cost is transferred from algorithm design to data acquisition, which reduces the difficulty of technology application. However, for cattle image segmentation, the complex breeding environment will be a problem. The color and texture of environmental information in the image will have an impact on the segmentation of cattle's details. Especially when FCN uses deconvolution operation in the process of up-sampling, it is insensitive to the details of the image and does not take into account the class relationship between the pixels, which makes the segmentation result lack of spatial regularity and spatial consistency, so the segmentation effect will be very rough. In order to improve the accuracy of semantics segmentation for fully convolutional networks and segmentation effect of cattle image details, this paper proposes a method of fully convolutional networks semantic segmentation based on RGBD cattle image. We create a concept which named "depth density". The value of depth density can quantify the probability about whether different pixels have the same category. According to the mapping relationship between RGB image and depth image on pixel level content, we optimize the semantic segmentation results of cattle's image by FCN. The experimental results showed that, better than FCN-8s, the proposed method could improve the pixel accuracy, mean accuracy, mean intersection over union and frequency weight intersection over union by 2.5%, 2.3%, 3.4% and 2.7% respectively.

Keywords: image processing; models; animals; semantic segmentation; RGB-D; fully convolutional networks; multimodal; cattle's image