基于三维重建的结构健康监测

胡芳侨

院 （系）：土木工程学院　　专　　业：土木工程力学精英班

学　　号：1143310207　　指导教师：李惠

2018 年 **6** 月

哈爾濱工業大學

毕业设计（论文）

题　　目　基于三维重建的结构健康监测

专　　业　土木工程力学精英班

学　　号　1143310207

学　　生　胡芳侨

指 导 教 师　李惠

答 辩 日 期　2018 年 6 月 19 日

摘　要

本文的目的是利用三维重建算法，对土木工程结构进行三维重建，并实现视觉健康监测。本文利用图像数据和重建出的三维点云数据，对建筑等结构的破坏进行识别，包括墙体裂缝、某个构件脱落等。破坏的位置会反映在三维点云模型上，方便检修人员进行定位、检查和维修等。

关键词：结构健康监测；计算机视觉；三维重建；损伤识别

Abstract

The purpose of this paper is to implement 3D reconstruction algorithm to civil engineering structures, achieving visual health monitoring of a structure. In this paper, the image data and the reconstructed 3D points data are used to detect the damage of buildings and other structures, including wall cracks, components fall off etc.. The positions are then reflected in the 3D point cloud model to facilitate professionals to further locate, detect and maintain this structure.

Keywords: Computer vision; Structural health monitoring; 3D reconstruction; Damage detection

目　录

摘　要 ..... **- 1 -**

ABSTRACT ..... **- 2 -**

第 1 章 绪　论 ..... **- 5 -**

1.1 课题背景及研究的目的和意义 ..... - 5 -

1.2 国内外在该方向的研究现状及分析 ..... - 5 -

1.2.1 利用无人机进行视觉健康监测的发展 ..... - 5 -

1.2.2 三维重建算法的发展 ..... - 6 -

1.2.3 基于图像的损伤识别和分割算法的发展 ..... - 6 -

1.2.4 点云处理算法的发展 ..... - 6 -

1.3 本文的主要研究内容 ..... - 6 -

第 2 章 基于图像的三维重建 ..... **- 8 -**

2.1 三维重建算法概述 ..... - 8 -

2.1.1 相机模型和相机参数 ..... - 8 -

2.1.2 特征点提取和特征点匹配 ..... - 9 -

2.1.3 计算基础矩阵 F 和外参数矩阵[R|T] ..... - 9 -

2.1.4 三角定位(Triangulation) ..... - 10 -

2.1.5 光束法平差(Bundle Adjustment) ..... - 11 -

2.1.6 多视图立体视觉(Multi-View Stereo) ..... - 11 -

2.2 北盘江大桥的三维重建 ..... - 11 -

2.2.1 数据说明 ..... - 11 -

2.2.2 特征点提取和匹配 ..... - 12 -

2.2.3 三角重建和 BA 优化 ..... - 12 -

2.2.4 多视图立体视觉(Multi-View Stereo) ..... - 13 -

2.3 某建筑的三维重建 ..... - 14 -

2.3.1 数据说明 ..... - 14 -

2.3.2 特征点提取和匹配 ..... - 15 -

2.3.3 三角重建和 BA 优化 - 15 -
2.3.4 多视图立体视觉(Multi-View Stereo) - 16 -
第 3 章 基于图像的损伤识别 - 18 -
3.1 图像的语义分割/实例分割 - 18 -
3.1.1 U-Net 网络架构 - 18 -
3.1.2 训练过程 - 19 -
3.2 某建筑的损伤识别 - 20 -
3.2.1 训练集数据说明 - 20 -
3.2.2 测试集表现 - 21 -
第 4 章 点云处理 - 23 -
4.1 点云匹配算法概述 - 23 -
4.2 点云比较算法概述 - 23 -
4.3 某建筑的点云匹配 - 25 -
4.4 某建筑的点云比较 - 25 -
4.5 某建筑的局部裂缝的三维重建和登记 - 28 -
结 论 - 30 -
参考文献 - 31 -
哈尔滨工业大学本科毕业设计（论文）原创性声明 - 33 -
致 谢 - 34 -
附录I 文献翻译原文 - 35 -
附录II 文献翻译译文 - 43 -

第 1 章 绪　论

1.1 课题背景及研究的目的和意义

结构的损伤和老化问题是一个重要的问题，结构健康监测和剩余寿命评估是一个必不可少的工作，并且在接下来的几十年中将变得越来越重要。视觉健康监测作为一种无损的健康监测方式，经常用于桥梁等结构的日常检查和维护，虽然视觉监测很大程度上依赖于专家经验。在大型桥梁、超高层建筑等定期进行人工视觉监测和报告是一个费时费力费财的工作，例如最常用的桥下监测单元和爬索缆车等工具，它们体积大、造价高。随着无人机（UAV）技术的成熟，无人机拍摄视频和照片变得越来越容易，我们有了一个新的途径解决这些问题。借助无人机拍摄的图像和视频，我们可以得到很多有用的信息。

最新发展的计算机视觉和图像处理技术提供了一种自动的监测手段，可以捕捉局部的损伤，既不需要昂贵的传感器，也更少依赖于专家经验。目前无人机在健康监测中的作用主要是拍摄图像和视频，对于这些数据，主流的方法是利用这些数据，然后基于计算机视觉的识别算法如物体检测、语义/实例分割等，来检测裂缝等损伤。需要注意的是这些方法都没有利用结构和相机的空间特征。三维重建很好地解决了这一点，利用三维重建算法可以同时恢复出结构的空间位置和相机的空间位置，利用得到的三维点云数据，可以得到更多的信息如局部位移变化、空间上复杂的损伤信息，并且有利于将整个结构的损伤信息进行整合，将损伤位置进行可视化表达。

该课题的意义在于 1) 可以减少甚至代替人工检测，节约人力物力；2) 结合了结构的三维信息，可以进行更多种类的损伤识别；3) 将整个结构的损伤情况整合到同一个三维模型中，并且实现可视化，有助于维修和管理。

1.2 国内外在该方向的研究现状及分析

该课题涉及的相关内容如下：利用无人机进行结构的视觉健康监测；通用的三维重建算法；重建后点云的分析如 3D 物体分割、点云匹配、点云比较等算法；基于图像的损伤识别，包括图像分割等算法；

1.2.1 利用无人机进行视觉健康监测的发展

利用无人机对桥梁等结构进行监测，国外有很多研究。整个视觉监测系统方面，有基于 UAV 的桥梁监测系统[1]；具体的方面，有基于无人机的半自动损伤监

测方法[2]，有利用无人机图像识别并监测裂缝的算法[3]，有利用照片的地理参照来建立三维点云模型，将监测结果和结构位置关联起来[4]等。

1.2.2 三维重建算法的发展

传统的基于图像进行三维重建的算法中，最适合本文的方法是运动相机恢复(Structure-from-Motion) [5]方法中的增量重建(Incremental SFM)。SFM 是一种经典的三维重建方法，它的原理是从照片中检测特征点，并对不同视角的照片进行匹配，经典的方法为 SIFT[6]方法。匹配完成后，将匹配得到的点矩阵(Measurements)分解为相机运动矩阵(Motion)和结构矩阵(Structure)，借助基础矩阵(Fundamental Matrix)矩阵消除任意性，最后进行 光束法平差(Bundle Adjustments)优化。目前最好的 SFM 框架是 COLMAP[7]。结合 MVS[8]等算法，可得到稠密点云。

1.2.3 基于图像的损伤识别和分割算法的发展

最近的一项研究对土木工程中对基于图像或视频的裂缝识别进行了调查和评估，包括基于视觉的裂缝检测技术、用于桥梁等结构局部损伤检测等。

传统的计算机视觉的方法包括图像去噪(滤波)、边缘检测(一般用 Canny 算子)、直线检测(Hougn 变换)、形态学函数、颜色成分分析、纹理检测、小波变换，聚类和模式识别等。

和机器学习结合的损伤识别算法大致有以下几种：生成对抗网络(GAN)；卷积神经网络(CNN)；最优熵阈值方法；种子区域生长算法；边缘检测(基于深度学习的)算法等。

图像分割领域流行的算法主要有三类：1) 全卷积神经网络(FCN)进行分类和条件随机场网络(CRF)进行优化，以 Deeplab V3+Chen, Zhu [9]，U-Net[10]等为代表；2) 递归神经网络(RNN)进行推理建模和条件随机场(CRF)进行优化；3)生成对抗网络(GAN)。

1.2.4 点云处理算法的发展

利用点云处理算法可以对物体进行匹配、比较、识别和分割。经典的点云匹配算法如 ICP[11]等；单个物体识别算法如 3D ShapeNets、VoxNet、PointCNN 等；多物体识别算法如[12]等；场景/物体分割算法如[13]等；

1.3 本文的主要研究内容

本课题的研究内容主要是利用三维重建算法，对土木工程结构进行三维重建，

以此实现该结构的视觉健康监测。首先利用航拍视频进行三维重建，其次利用视频帧和三维点云数据，对该结构的表面损伤进行识别，最后将损伤的位置反映在三维模型中。

第 2 章 基于图像的三维重建

2.1 三维重建算法概述

传统的通用的三维重建算法选取 Structure-from-Motion(SFM)算法，SFM 是由多幅同一场景不同视角的图像来估计场景的三维信息和相机位置的算法。它输入是同一对象的一组存在重叠视角的图像，从不同的角度拍摄，输出是对物体的三维重建，以及所有相机内在和外在的摄像机参数。通常情况下，从运动系统将这个过程分为三个阶段：特征检测与提取；特征匹配和几何验证；结构和运动重建。图 1 展示了三维重建算法的流程。

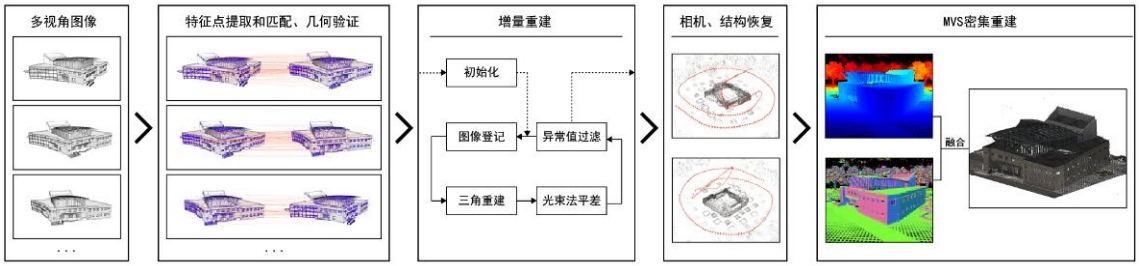

图 1：三维重建算法流程

2.1.1 相机模型和相机参数

2.1.1.1 像平面、虚拟像平面和焦距 f

以常用的小孔相机模型为例，如图 2(左)所示。Film 平面通常被称为图像平面（像平面）或视网膜平面，中间的小孔为针孔或相机中心。像平面和小孔 O 之间的距离为焦距 f。为了表达更简单，在计算机视觉的常用另一种表达方式，像平面被放置在 O 与对象之间，在这种情况下，它被称为虚拟像平面。与小孔相机类似，透镜相机如图 2(右)所示。焦距 f 联系了相机坐标系中一点的位置和像平面一点的位置。对于透镜相机，利用相似三角形的关系可以得 p'= z'p/z 。

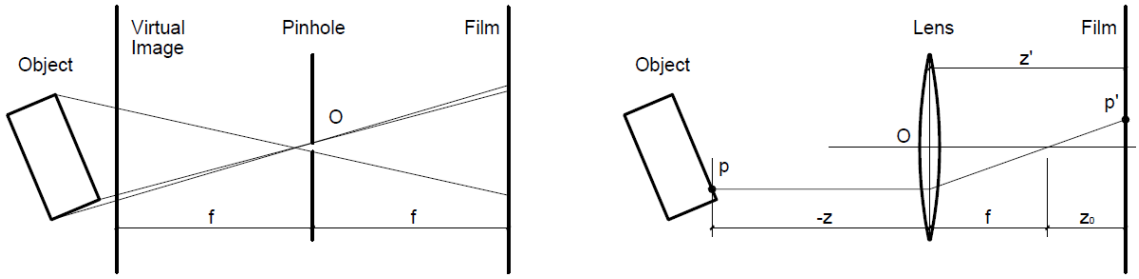

图 2：小孔相机模型(左)和镜头相机模型(右)

2.1.1.2 坐标系和坐标变换

世界坐标系为真实世界的坐标系，在其中的点 P 以(x, y, z)表示，在一次重建过程中可以以第一个相机的坐标系为世界坐标系；其中相机坐标系以镜头中心为原点，以沿光轴向外为 z 方向，采用右手系；像平面坐标系中，在其上的点 p 以(u, v)表示，单位为像素。从真实世界中的某个对象到像平面上的像素图像，需要经历两次变换，即从世界坐标系到相机坐标系、从相机坐标系到像平面坐标系。前者为刚体变换，后者为投影变换。

2.1.1.3 畸变参数矩阵、内参数矩阵和外参数矩阵

对于透镜需要引入径向畸变参数 λ。$\lambda=1\pm\sum_{p=1}^{3}k_p d^{2p}$，其中 $d^2=au^2+bv^2+cuv$。现假设真实世界中的某对象上一点 P，经变换后落在像平面上一点 p，其变换关系为：

$$p=S_\lambda K[R|T]P \tag{1}$$

其中 $S_\lambda=\begin{bmatrix}1/\lambda & 0 & 0\\ 0 & 1/\lambda & 0\\ 0 & 0 & 1\end{bmatrix}$ 为相机镜头径向畸变参数矩阵；$K=\begin{bmatrix}\alpha & -\alpha\cot\theta & u_0 & 0\\ 0 & \beta/\sin\theta & v_0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$ 为相机内参数矩阵，它表示了从相机坐标系到像平面坐标系的投影变换关系；$[R|T]$为外参数矩阵，它表示了从世界坐标系到相机坐标系的刚体变换关系，即旋转矩阵 R 和平移向量 T。

2.1.2 特征点提取和特征点匹配

Scale-invariant feature transform(SIFT) 是一种特征点提取和匹配的算法，它将一张图像中的所有特征点用一个描述子来描述，相当于给每个特征点做一个不同的“标记”。其中描述子包含了尺度(Scale)和方向(Orientation)两个信息。对于两张图像，利用 SIFT 可以对带有描述子的特征点进行匹配。特征点的提取和匹配如图 3 所示。

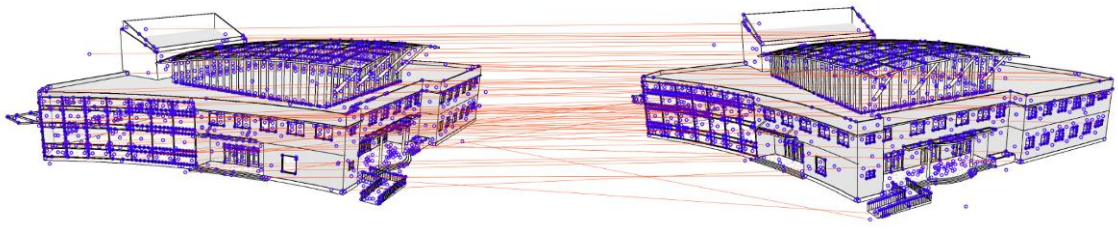

图 3：SIFT 特征点提取和匹配

2.1.3 计算基础矩阵 F 和外参数矩阵[R|T]

2.1.3.1 本征矩阵 E 和基础矩阵 F

三角重建过程中有两个重要的矩阵：本征矩阵 E (Essential Matrix)和基础矩阵 F (Fundamental matrix)。它们的形象展示如图 4 所示。

本征矩阵 E=$[T_x]R$，它包含了一个相机相对另一个相机的刚体变换关系，其中 p_l 和 p_r 分别表示左相机图像和右相机图像的对应点。本征矩阵 E 将第一个相机得到的图像中的观测到的点 p_l 和第二个相机得到的图像中的点 p_r 关联起来。两者满足如下关系： $p_l[T_x]Rp_r=0$。

如果考虑相机的内参数，则需要引入基础矩阵。基础矩阵 $F=K_l^{-T}[T_x]RK_r^{-1}$，它也将第一个相机得到的图像中的观测到的点 p_l 和第二个相机得到的图像中的点 p_r 关联起来。两者满足如下关系：

$$p_l^T K_l^{-T}[T_x]RK_r^{-1}p_r^T=0 \tag{2}$$

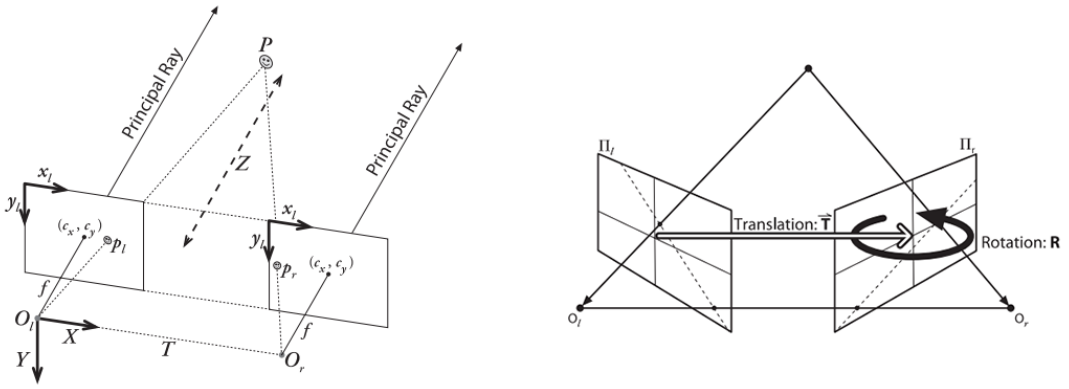

图 4：本征矩阵和基础矩阵

2.1.3.2 八点法计算基础矩阵 F

一般使用八点法来确定 F，即找到 8 个以上的对应点。对每一个点对，均有 $p_l F p_r=0$，以向量表示为 (W,f)=0，其中 $W=p_l^T p_r$，f 为 F 展开的向量，解出 $\hat{F}$。即 F 满足 $det(F)=0$ 和 $\min \| F-\hat{F} \|$。

2.1.3.3 计算外参数矩阵[R|T]

标定相机内参数或估计相机内参数(后者较为常用)后，得到内参数矩阵 K，可利用 $E=K^T FK$ 计算本征矩阵 E，进而可得到外参数矩阵[R|T]。

2.1.4 三角定位(Triangulation)

三角定位(Triangulation)的目的是利用已经得到的特征点对 p_l 和 p_r，借助内参数 K 和外参数[R|T]来恢复出世界坐标系中该点 P。

$$\min d(p, M_1 P') + d(p', M_2 P') \tag{3}$$

其中 $M = K[R|T]$。

2.1.5 光束法平差(Bundle Adjustment)

光束法平差(Bundle Adjustment)的作用是作为 SFM 的最后一步优化。它利用计算得到的结构三维点的坐标和计算、估计得到的相机内外参数矩阵进行重投影，以重投影误差最小为迭代目标。重投影误差的来源为图像噪点、相机内外参数矩阵估计误差等，它们使计算得到的值和实际情况不会完全相符。目前常用的是 Levenberg-Marquardt 迭代算法，使目标函数最小。目标函数为：

$$\mathrm{E}(M,X)=\sum_{i=1}^{m}\sum_{j=1}^{n}D(x_{ij},M_iX_j)^2 \tag{4}$$

其中 D 为非线性映射，x_{ij} 为特征点对，M_iX_j 为二次投影，m 为图像的个数，n 为两张图像匹配上的特征点个数。

2.1.6 多视图立体视觉(Multi-View Stereo)

多视图立体视觉（MVS）以 SFM 的输出作为输入，利用特征点的三维坐标和相机的坐标、视角，计算每张图像中每个像素的深度和法线信息。最后将每张图像的深度图和法线图融合，产生一个密集的点云。

2.2 北盘江大桥的三维重建

北盘江大桥位于贵州和云南的交界处，它属于斜拉桥。其大致数据：主跨 720m，主桥 1232m，引桥 68m，全桥 1341.4m。主桁架高 8m，斜拉索用钢绞线，桥塔采用 H 型索塔，塔高分别为为 269m(贵州一侧)和 247m(云南一侧)。

2.2.1 数据说明

本算法采用的数据为北盘江大桥环桥航拍视频。其中相机参数未知，得到了三段视频，它们的参数如下：分辨率为 4096*2160，帧速率为 25 帧/秒，视频总长度为 19 分 30 秒，比特率为 60131kbps。对视频进行按 FFmpeg 运动幅度帧提取，得到共 115 张图像。见图 5。

图 5：桥梁三维重建采用的 115 张图像

2.2.2 特征点提取和匹配

特征提取采用的相机模型为简单径向(Simple Radial)模型，包含 4 个参数：焦距 f，主点坐标(cx,cy)，第一个径向畸变参数 k，估计值为[4915.2,2048,1080,0.000]。SIFT 算法中关键参数如下，选取 Octaves 个数为 4，Octave 分辨率为 3，峰值阈值 0.00667，边缘阈值 10，最小、最大尺度限制比例分别为为 0.1667、3.0000。见图 6。

图 6：对所有图像进行特征点提取

特征点匹配策略采用详尽匹配(匹配每个可能像对)，并采取交叉验证，最大图像旋转限制比例为 0.8，最大图像距离限制比例为 0.7，最大容许误差为 4px。见图 7。

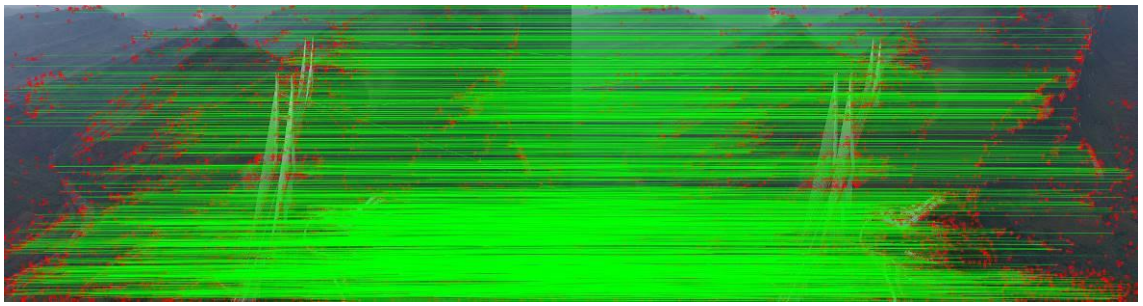

图 7：特征点匹配

2.2.3 三角重建和 BA 优化

关键参数如下，不选取初始匹配对，初始化匹配对最大误差限制为 4.00，最小

视角差为 16 度；新图像登记最大容许误差为 12px；三角重建最大角度误差限制为 2 度，最大重投影误差为 4px；由于事先没有用相机内参数，采用的是估计值，故 BA 优化同时优化相机参数，包括主点坐标(cx,cy)。所有 115 张图片成功登记并重建，得到 115 个相机位姿和 44907 个特征点。所得到的特征点云和相机位姿如图 8 所示。

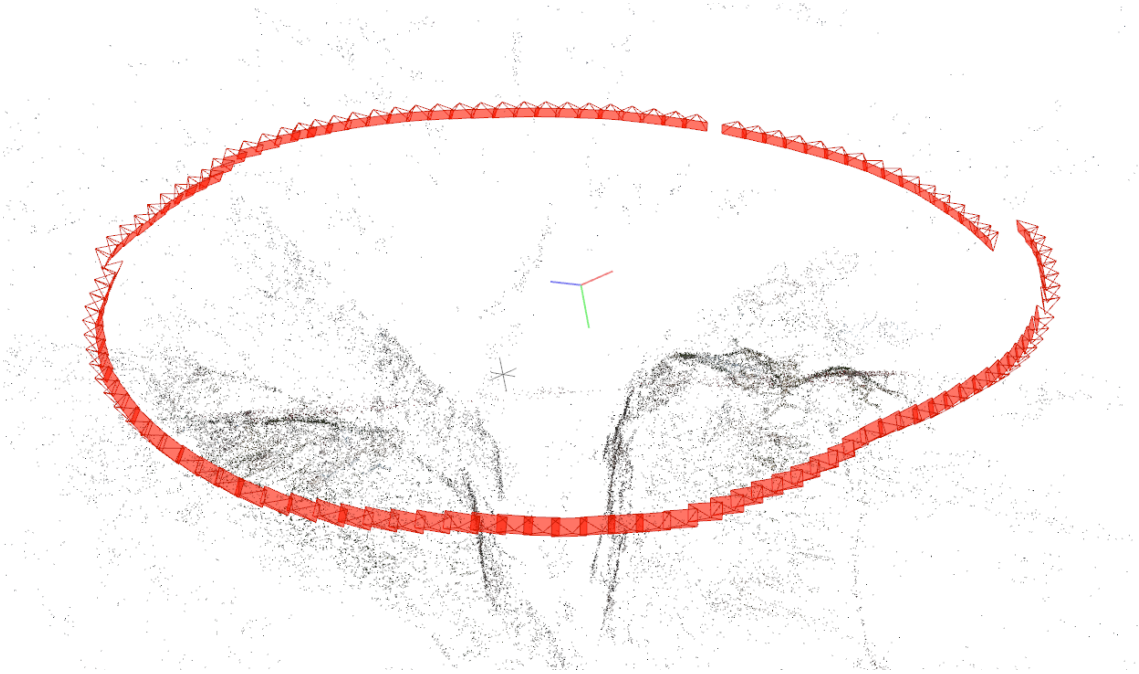

图 8：特征点云和相机位姿

2.2.4 多视图立体视觉(Multi-View Stereo)

从原始图像得到每个像素的深度信息和法线信息，从而得到深度图和法线图，然后进行点云融合。以一张图像举例，图 9 所示三张图分别为原始 RGB 图像，深度图，法线图。

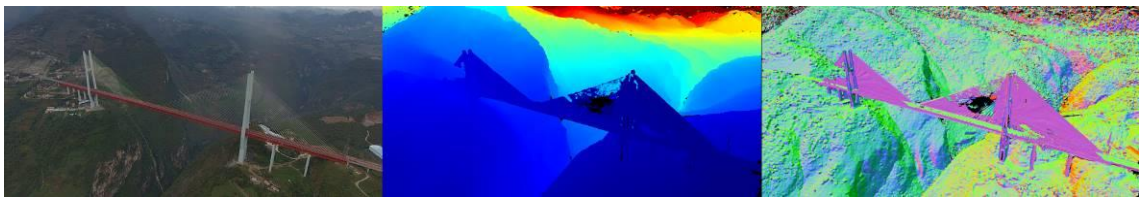

图 9：原始图像、深度图和法线图

最终融合的点云结果如图 10 所示。

图 10：融合后的密集点云

2.3 某建筑的三维重建

由于没有桥梁破坏后的数据，为验证算法的可行性，本文先从某建筑入手，对该建筑破坏前和破坏后分别进行三维重建。但是由于同样没有现实数据，故采用计算机来模拟一栋场馆建筑破坏前和破坏后的三维模型，并设置摄像机来模拟无人机航拍。

2.3.1 数据说明

本算法采用的数据来源如下，对某场馆建筑物进行三维建模，然后模拟破坏效果。破坏效果包括雨棚角度的改变、屋顶支撑杆的破坏、曲面屋顶的消失、门厅柱子的破坏、门窗的消失、墙面裂缝和墙角大裂缝等。对初始状态和破坏后的建筑分别进行视频渲染。相机参数如下：DLSR 传感器，对角线尺寸为 43mm，长宽比为 3:2，焦距为 27.9mm，视野角度 60 度，曝光模式自动，无运动模糊。得到了两段渲染视频，它们的参数如下：分辨率为 1920*1280，帧速率为 30 帧/秒，长度为 7 秒，比特率为 50617kbps(未压缩)。对两段视频进行所有帧提取，得到两组图像，每组 224 张。见图 11。

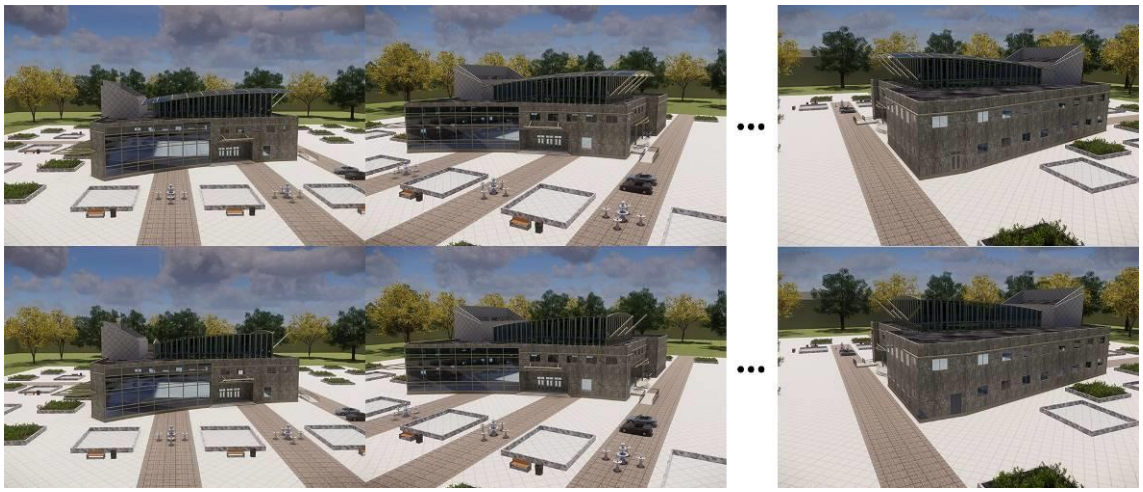

图 11：建筑三维重建采用的 448 张图像

2.3.2 特征点提取和匹配

特征提取采用的相机模型为简单径向(Simple Radial)模型，包含 4 个参数：焦距 f，主点坐标(cx,cy)，第一个径向畸变参数 k，估计值为[2400,960,640,0.000]。SIFT 算法中关键参数如下，选取 Octaves 个数为 4，Octave 分辨率为 3，峰值阈值 0.00667，边缘阈值 10，最小尺度限制为 0.1667，最大尺度限制为 3.0000(比例)。

特征点匹配策略采用详尽匹配(匹配每个可能像对)，并采取交叉验证，最大旋转限制为 0.8，最大距离限制为 0.7(比例)，最大容许误差为 4px。

2.3.3 三角重建和 BA 优化

关键参数如下，不选取初始匹配对，初始化匹配对最大误差限制为 4.00，最小视角差为 16 度；新图像登记最大容许误差为 12px；三角重建最大角度误差限制为 2 度，最大重投影误差为 4px；由于事先没有用相机内参数，采用的是估计值，故 BA 优化同时优化相机参数，包括主点坐标(cx,cy)。所有 224 张图片成功登记并重建，得到 224 个相机位姿和 102227 个特征点。所得到的特征点云和相机位姿如图 12 所示。

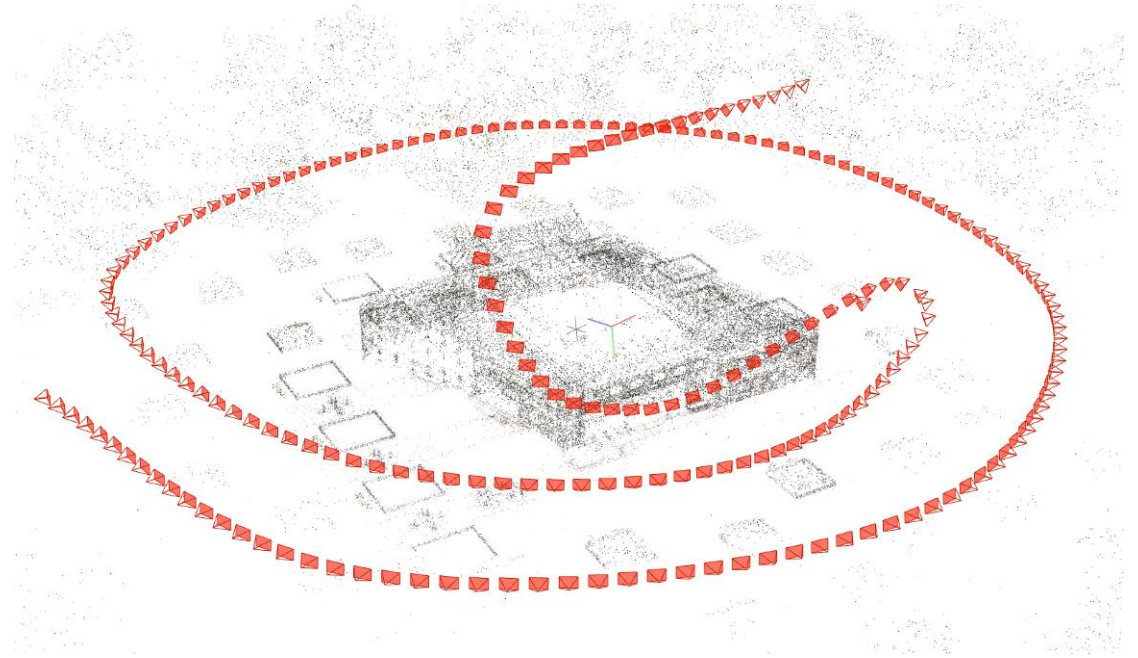

图 12：特征点云和相机位姿

2.3.4 多视图立体视觉(Multi-View Stereo)

从原始图像得到每个像素的深度信息和法线信息，从而得到深度图和法线图，然后进行点云融合。以一张图像举例，图 13 所示三张图分别为原始 RGB 图像，深度图，法线图。

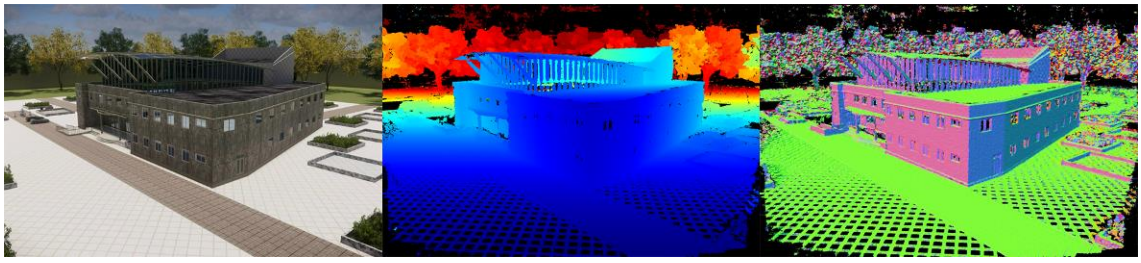

图 13：原始图像、深度图和法线图

最终融合的点云结果如图 14 所示。

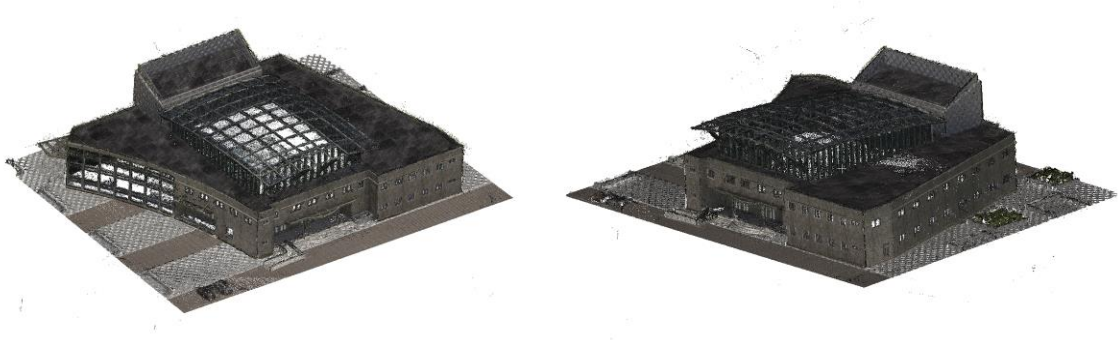

图 14：融合后的密集点云

第 3 章 基于图像的损伤识别

3.1 图像的语义分割/实例分割

语义分割(Semantic Segmentation)不同于图像分类，前者是对一张图像做一个分类，后者是对一张图片中的每一个像素做分类。实例分割(Instance Segmentation)的目标是对图像中每个实例进行单独分割。例如当图像中有两个人，语义分割会将两个人的像素作为一个分类，而实例分割会将两个人的像素分别做一个分类。语义分割的数据标注更简单，对于简单的裂缝识别可考虑语义分割。

目前图像分割流行的算法主要有三类：1) 全卷积神经网络(FCN)进行分类和条件随机场网络(CRF)进行优化；2) 递归神经网络(RNN)进行推理建模和条件随机场(CRF)进行优化；3)生成对抗网络(GAN)。本文采用第一类中的 U-Net 网络。

3.1.1 U-Net 网络架构

网络的结构包含了编码路线(左半部分)和解码路线(右半部分)。见图 15。

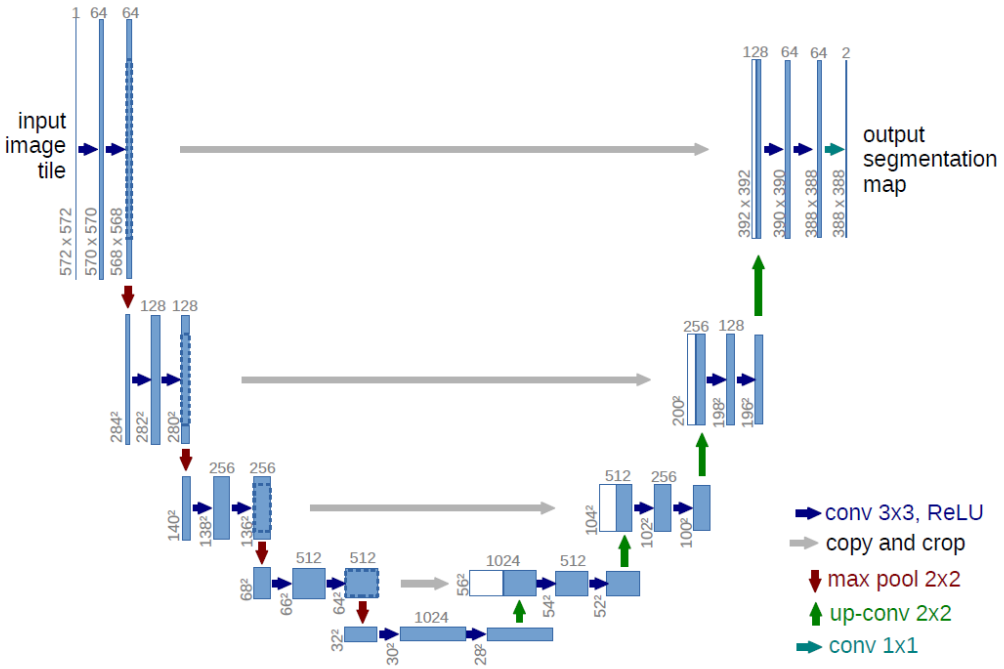

图 15：U-Net 网络架构

蓝色层为卷积层(Convolutional Layer)和激活层(Activation Layer)，在图像领

域，卷积操作是利用特定的卷积核在图像上按一定的步长滑动，将图像上卷积核所对应的位置某个通道的像素值和卷积核进行乘和加的操作。一系列卷积操作可以将图像的高维特征图提取出来，该网络采用卷积核大小为 3×3。由于卷积操作是一种线性运算，而图像大部分特征用线性运算不足以描述，所以需要加入非线性激活层。激活函数的特点是处处可微、输入区间为任意、输出区间为[0, 1]。该网络采用的激活函数为 ReLU。

红色箭头代表最大值池化(Max Pool)操作，池化为下采样操作，最大值池化对核大小的区域内选取最大的特征值保留下来。它可以保留图像的最重要的特征。该网络采用的是 2×2 Max Pooling。

绿色箭头代表上卷积(Up Convolutional)操作。上卷积操作为一个上采样的过程，可以理解为特征图的放大，并且通道数减半。上卷积的一般做法是在被卷积图像的像素之间插入 0(空白像素)，做放大处理，然后在这个放大的特征图上进行卷积。

灰色的箭头代表将特征图进行复制、裁剪。低维特征图(前部分的特征图)的特点是对像素的定位效果比较好；而高维特征图的特点是对像素的分类效果比较好。复制和融合的操作可以将网络开始的低维特征图利用起来，避免损失位置信息，一定程度上减小了“特征图放大”这个过程对像素级分割的定位造成的不利影响。

最后层的绿色的箭头为 1×1 的卷积操作，它将 64 通道的特征图映射为想要得到的类别个数通道的图像。

3.1.2 训练过程

该网络的输入为一系列原图像和它们的分类标签遮罩图像。利用带动量的随机梯度下降法(SGD)训练。

$$w := w - \eta \nabla Q_i(w) + \alpha \Delta w \tag{5}$$

其中η 为学习率，α 为混合权重。相比较梯度下降法，随机梯度下降法每次随机使用一个或一个批次的数据量来计算梯度并更新参数。随机梯度下降法的优点有两点，首先它不需要计算所有数据来计算梯度，在图像领域利用所有数据计算梯度也会造成很大的计算开销；其次它随机选取数据，可能会使优化过程跳出局部极小值。而带动量的随机梯度下降法将本次迭代更新后的权重和上一次的做一个加权混合，这样可以避免优化过程中产生的震荡，使优化过程更快收敛。它们的对比如图 16 所示。

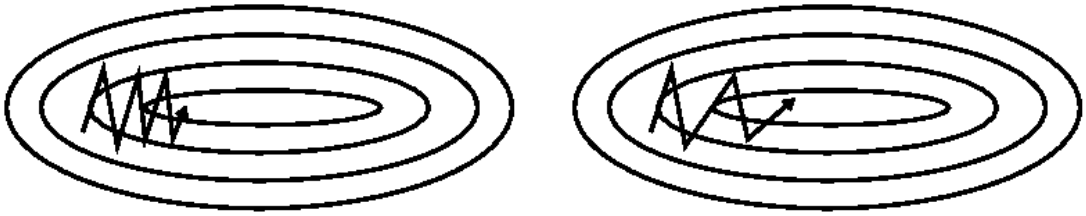

图 16：不带动量的梯度下降法(左)和带动量的梯度下降法(右)

利用最后一层特征图上的像素级的 Soft-max 和交叉熵(Cross Entropy)损失函数的结合，来定义能量函数。Soft-max 的定义为：

$$p_k(x)=\exp(a_k(x))/(\sum_{k'=1}^{K}\exp(a_{k'}(x))) \tag{6}$$

其中 $a_k(x)$ 为激活函数，在第 k 个特征图通道，像素位置为 x 处；K 为分类的个数；$p_k(x)\approx 1$ 表示该通道的特征图为该类别的概率是最大的。交叉熵对每个位置的 $p_{l(x)}(x)$ 距离 1 的偏差做一个惩罚：

$$E=\sum_{x\in\Omega} w(x)\log(p_{l(x)}(x)) \tag{7}$$

其中 $l:\Omega\to\{1,\ldots,K\}$ 为每个像素的真实标签，$w:\Omega\to\mathbb{R}$ 是一个权重图，利用这个权重图，在训练时可以给某些像素更高的权重。分割边界用形态学操作来计算，权重图的计算如下：

$$w(x)=w_c(x)+w_0\cdot\exp(\frac{-d_1(x)+d_2(x)^2}{2\sigma^2}) \tag{8}$$

其中 $w_c:\Omega\to\mathbb{R}$ 用来平衡类之间频率的权重图，$d_1:\Omega\to\mathbb{R}$ 表示到最近细胞边界的距离，$d_2:\Omega\to\mathbb{R}$ 表示离第二近的。设定 $w_0=10$ ，$\sigma\approx 5px$ 。

一个好的初始化权重参数对网络至关重要，否则可能导致网络的某些部分过于活跃，而其它部分只做出了很少的贡献。初始的权重参数采用高斯分布，标准差为 $\sqrt{2/N}$ ，其中 N 为该神经单元的输入节点的个数。对于一个 64 通道的特征图，3×3 的卷积核，$N=9\times 64=576$。

3.2 某建筑的损伤识别

3.2.1 训练集数据说明

输入的训练数据为 178 张包含裂缝的图像和裂缝的遮罩图像。见图 17。原始图像的尺寸为 1920*1280*3，为 RGB 模式；遮罩图像的尺寸为 1920*1280*1，为二值化图像。这些数据中，裂缝均为同一裂缝，而且光照条件恒定，背景也比较

相似，即数据的分布比较集中，很容易导致过拟合问题。但是人工标注数据是一个非常消耗时间和精力的工作，并且需要多人来分工完成，所以仅仅是以实现算法为目标。

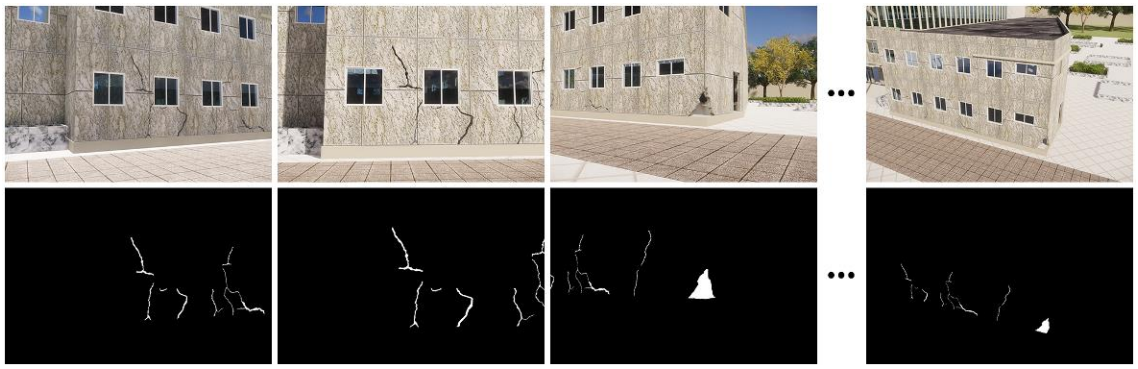

图 17：训练集数据示例

3.2.2 测试集表现

测试集为多张包含裂缝的原图像。相比较训练集，改变了裂缝的形状、光照条件、建筑的立面、背景等条件，甚至使用了其它的场景。得到的一些结果如图 18 所示。

图 18：测试集结果示例

由此可见，在该训练集上 U-Net 的确学习到了裂缝尤其是细小裂缝的一些特征，并且对光照变化、背景变化等有一定的免疫能力。但由于训练集数据分布过于集中，导致训练得到的模型的泛化能力不够强。

第 4 章 点云处理

对点云的处理主要为以下两方面：1) 局部点云模型到整体点云模型的匹配；2) 两个点云模型的差异比较和可视化。

4.1 点云匹配算法概述

截止目前，对两个点云进行自动匹配的算法主要是 Iterative Closest Point (ICP)算法。ICP 算法能够找到一个刚体变换矩阵，使同一目标的两个不同坐标系下的点云匹配到同一坐标系统中，即找到从一个坐标系到另一个坐标系的一个刚体变换矩阵[R|T]。注意当两个点云的 Scale 不一致时，需要添加一个缩放矩阵 S。见图 19。

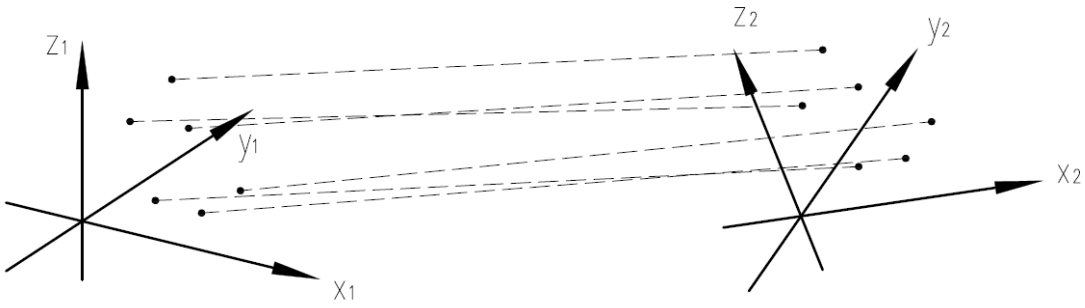

图 19：不同坐标系下的匹配

它实际上是基于最小二乘法的最优化。给定一个停止迭代的阈值，该算法重复计算两个点云的均方误差， 即计算刚体变换矩阵[R|T]的最优解，满足阈值则停止迭代。给定两个点集 X_1 和 X_2 (齐次坐标形式)，将这个问题描述为最小化均方误差(或者均方根误差)：

$$\mathrm{E}(X_1, X_2) = \sum_{i=1}^{m} ([R|T]X_1 - X_2)^2 \tag{9}$$

其中 m 为 X_1 中点的个数，通常情况选择更密集的 X_2 作为参照系。

ICP 法分以下几个步骤：1，对 X_1 中的每个点，在 X_2 中寻找最近点(欧氏距离)；2，对 X_1 加以刚体变换，使得上式最小；3，开始迭代，直到满足停止迭代的条件，这个条件包括迭代次数和误差阈值。

4.2 点云比较算法概述

点云比较整体采用最近点(KNN)策略。对于两个点云，选取一个作为参照(通常是更密集的)，称为参照云；一个做比较，称为比较云。对于比较云的每一点，搜索参考云中最近的点，并计算它们的欧氏距离。

点到点的距离因为误差较大，故目前比较常用的是在参照云局部做一个表面，来计算比较云中的点到参考云中的面的距离。具体做法是通过在“最近”点和它的 K 个临近点(KNN)上使用一个数学模型来模拟参考云局部表面。对于平面占大多数的物体通常选用最小二乘平面(Least Square Plane)或三角形平面(2D1/2 Delaunay Triangulation)，对于多曲面物体则采用二次曲面(Quadric)。

如图 20 所示，∑为参照云局部某区域，P 为比较云上一点。P 到∑的距离一般用 P 到∑局部表面模型来表示。

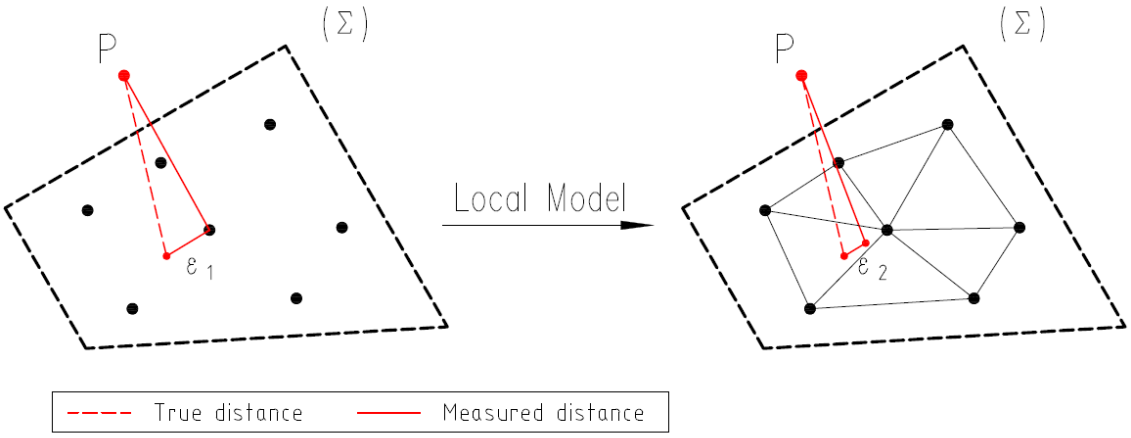

图 20：点云比较

采用八叉树(Octree)来对点云进行细分，八叉树可以看作对三维空间的二分法。细分水平越高，八叉树细胞就越小，每个单元格的点就越少，为了找到最近的一个(K 个)点，需要做的计算就越少。但反过来说，细胞越小，细胞需要被迭代搜索的次数越多，如果比较点离它最近的参考点很远，那么搜索时间会很长。所以大的点云需要更高八叉树层数，但是如果比较云的点离参考云很远，那么低的八叉树层数会更好。八叉树示例见图 21。

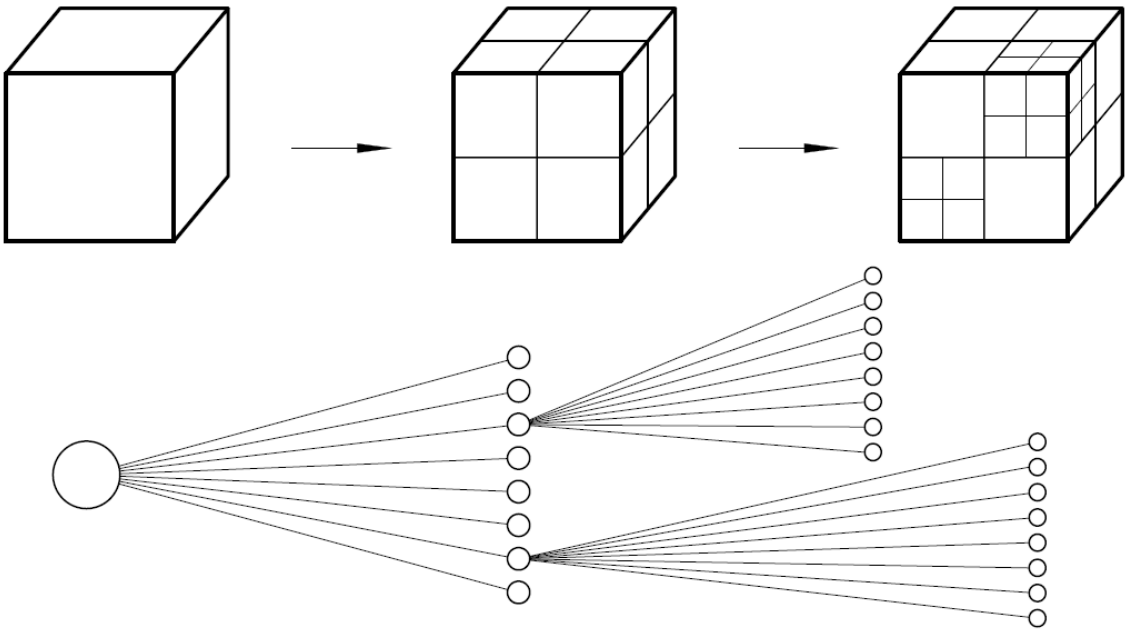

图 21：八叉树划分点云

4.3 某建筑的点云匹配

迭代停止的均方根(RMS)误差为 1e-24，同时调整了 Scale，得到的 $S[R|T]$ 矩阵(齐次坐标形式)为 $\begin{bmatrix} 0.990 & -0.035 & 0.183 & -0.020 \\ 0.023 & 1.005 & 0.070 & -0.036 \\ -0.185 & -0.065 & 0.988 & 0.004 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$，其中 Scale 值为 1.00751。匹配前后的结果见图 22。

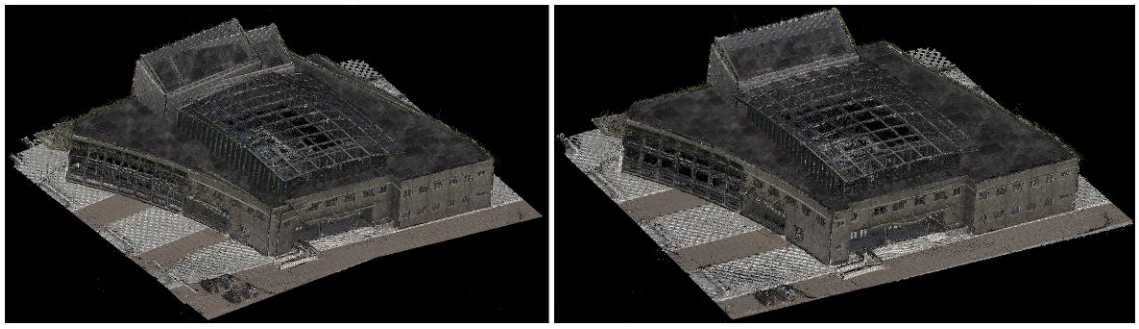

图 22：匹配前(左)和匹配后(右)

4.4 某建筑的点云比较

点云比较采用的八叉树层数为 8，局部平面模型采用最小二乘平面(Least Square

Plane)，最近点(KNN)个数为 6。

对于点云结果，较大尺度的破坏，如柱子破坏、门窗脱落等，点云对比可得到较好的结果。但对于裂缝等细节破坏，点云对比的结果过于粗糙，无法识别。点云比较的结果如图 23 所示。

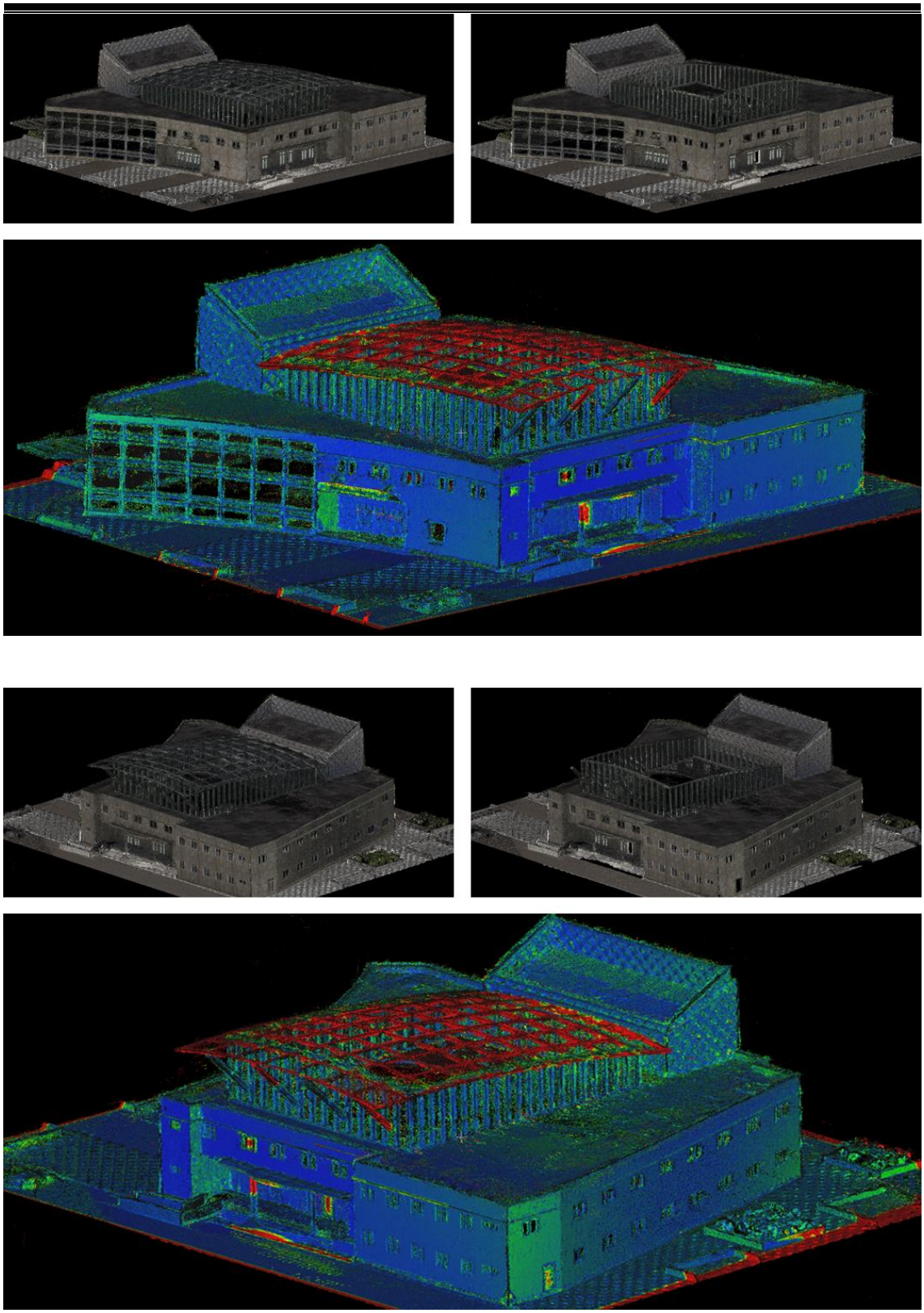

图 23：破坏前和破坏后的比较

4.5 某建筑的局部裂缝的三维重建和登记

对在图像上识别出的裂缝遮罩做 2px 膨胀处理，将遮罩叠加到原始图像上，混合权重为 0.5。对这些图像进行局部三维重建，得到的结果如图 24 所示。

图 24：局部裂缝的三维重建

先采用四点匹配法进行粗略匹配，得到的 $S[R|T]$ 矩阵为 $\begin{bmatrix} 0.130 & -0.013 & 0.065 & 1.225 \\ 0.007 & 0.145 & 0.015 & 0.823 \\ -0.066 & -0.010 & 0.129 & -1.554 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$，其中 Scale 值为 0.145675。进行精细匹配，迭代停止的均方根(RMS)误差为 1e-24，同时调整了 Scale，得到的 $S[R|T]$ 矩阵为 $\begin{bmatrix} 1.002 & -0.009 & 0.003 & -0.002 \\ 0.009 & 1.002 & -0.001 & -0.008 \\ -0.003 & 0.001 & 1.002 & 0.006 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$，其中 Scale 值为 1.00201。匹配结果见图 25。

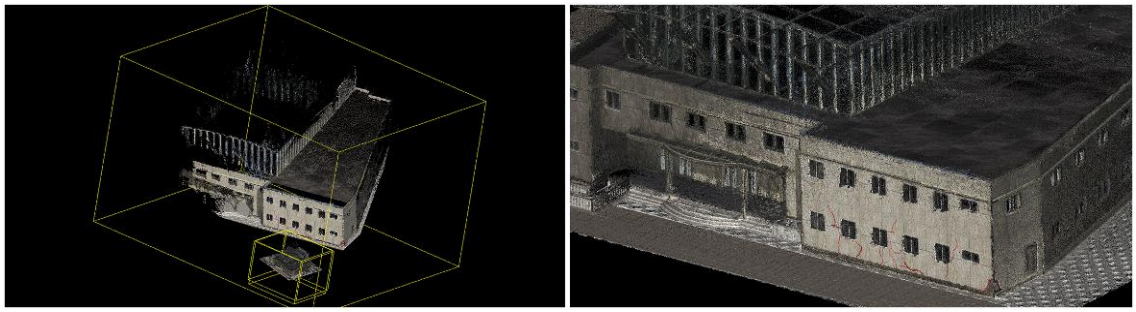

图 25：匹配前和匹配后

结　论

本文从三维重建的角度出发，讨论了结构的视觉健康监测的可行性。在三维重建的基础上，本文做到了破坏检测，并且将其可视化在三维中。本文的工作流程总结如下：

1) 对一个结构进行破坏前和破坏后的三维重建，三维重建所采用的的算法为SFM(运动相机恢复)和 MVS(多视图立体视觉)。

2) 对于大尺度的破坏，利用点云的对比可以得到较为直观的结果，点云匹配和对比采用的算法为 ICP(迭代最近点)和八叉树细分后的 KNN(最近点)等算法。

3) 对于细节部分的破坏，粗糙的点云对比无法得到较好的结果，本文采用的是像素级的图像语义分割，将图像中的裂缝等分割出来，然后再次利用三维重建算法(SFM 和 MVS)将局部的破坏重建出来，利用点云匹配(ICP)将局部三维点云匹配到整体中，达到可视化的目的。

但本文同样存在很多不足之处，未来也有一些工作需要完善：

1) 对于桥梁等自相似性明显的结构，传统的三维重建算法很难得到重建结果，特征点的匹配多依赖于地形和桥梁的明显纹理；即使得到局部的三维重建结果，也很难将其匹配到整体中，结合先验知识可能会解决这一问题。

2) 本文仅做到了将破坏检测出来，并得到它们在结构中的位置，但破坏的划分依据仅仅是点到面的距离，没有和实际情况结合。

3) 本文没有做到对破坏分类。

参考文献

1. Chan, B., et al., *Towards UAV-based bridge inspection systems: A review and an application perspective.* Structural Monitoring and Maintenance, 2015. **2**(3): p. 283-300.
2. Hallermann, N. and G. Morgenthal. *Visual inspection strategies for large bridges using Unmanned Aerial Vehicles (UAV)*. in *Proc. of 7th IABMAS, International Conference on Bridge Maintenance, Safety and Management*. 2014.
3. Yu, H., et al. *A UAV-based crack inspection system for concrete bridge monitoring*. in *Geoscience and Remote Sensing Symposium (IGARSS), 2017 IEEE International*. 2017. IEEE.
4. Hallermann, N. and G. Morgenthal. *Visual inspection strategies for large bridges using Unmanned Aerial Vehicles (UAV)*. in *Iabmas*. 2014.
5. Koenderink, J.J. and A.J. van Doorn, *Affine structure from motion.* Journal of the Optical Society of America A Optics & Image Science, 1991. **8**(2): p. 377-385.
6. Lowe, D.G., *Distinctive Image Features from Scale-Invariant Keypoints.* International Journal of Computer Vision, 2004. **60**(2): p. 91-110.
7. Schönberger, J.L. and J.M. Frahm. *Structure-from-Motion Revisited*. in *Computer Vision and Pattern Recognition*. 2016.
8. Seitz, S.M., et al. *A comparison and evaluation of multi-view stereo reconstruction algorithms*. in *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*. 2006. IEEE.
9. Chen, L.-C., et al., *Encoder-decoder with atrous separable convolution for semantic image segmentation.* arXiv preprint arXiv:1802.02611, 2018.
10. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.
11. Besl, P.J. and N.D. McKay. *Method for registration of 3-D shapes*. in *Sensor Fusion IV: Control Paradigms and Data Structures*. 1992. International Society for Optics and Photonics.

12. Qi, R.C., *Object Detection in 3D Scenes Using CNNs in Multi-view Images.* 2016.

13. Yi, L., et al., *Learning hierarchical shape segmentation and labeling from online repositories.* arXiv preprint arXiv:1705.01661, 2017.

哈尔滨工业大学本科毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学攻读学士学位期间，所提交的毕业设计（论文）《基于三维重建的结构健康监测》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名：　　日期：　　年　　月　　日

致　谢

衷心感谢导师李惠教授对本人的精心指导和帮助。在课题上她很为我选定了一个非常契合我的课题，并且深入到细节耐心指导我；在经济上，只要是对科研课题有帮助的，她都会尽力支持我。她的言传身教将使我终生受益。

感谢李惠教授，以及课题组全体老师和同窗们的热情帮助和支持！

附录Ⅰ 文献翻译原文

U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger, Philipp Fischer, and Thomas Brox
Computer Science Department and BIOSS Centre for Biological Signalling Studies,
University of Freiburg, Germany
ronneber@informatik.uni-freiburg.de,
WWW home page: http://lmb.informatik.uni-freiburg.de/

Abstract. There is large consent that successful training of deep networks requires many thousand annotated training samples. In this paper, we present a network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. We show that such a network can be trained end-to-end from very few images and outperforms the prior best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Using the same network trained on transmitted light microscopy images (phase contrast and DIC) we won the ISBI cell tracking challenge 2015 in these categories by a large margin. Moreover, the network is fast. Segmentation of a 512x512 image takes less than a second on a recent GPU. The full implementation (based on Caffe) and the trained networks are available at http://lmb.informatik.uni-freiburg.de/people/ronneber/u-net.

1 Introduction

In the last two years, deep convolutional networks have outperformed the state of the art in many visual recognition tasks, e.g. [7,3]. While convolutional networks have already existed for a long time [8], their success was limited due to the size of the available training sets and the size of the considered networks. The breakthrough by Krizhevsky et al. [7] was due to supervised training of a large network with 8 layers and millions of parameters on the ImageNet dataset with 1 million training images. Since then, even larger and deeper networks have been trained [12].

The typical use of convolutional networks is on classification tasks, where the output to an image is a single class label. However, in many visual tasks, especially in biomedical image processing, the desired output should include localization, i.e., a class label is supposed to be assigned to each pixel. Moreover, thousands of training images are usually beyond reach in biomedical tasks. Hence, Ciresan et al. [1] trained a network in a sliding-window setup to predict the class label of each pixel by providing a local region (patch) around that pixel as input. First, this network can localize. Secondly, the training data in terms of patches is much larger than the number of training images. The resulting network won the EM segmentation challenge at ISBI 2012 by a large margin.

Obviously, the strategy in Ciresan et al. [1] has two drawbacks. First, it is quite slow because the network must be run separately for each patch, and there is a lot of

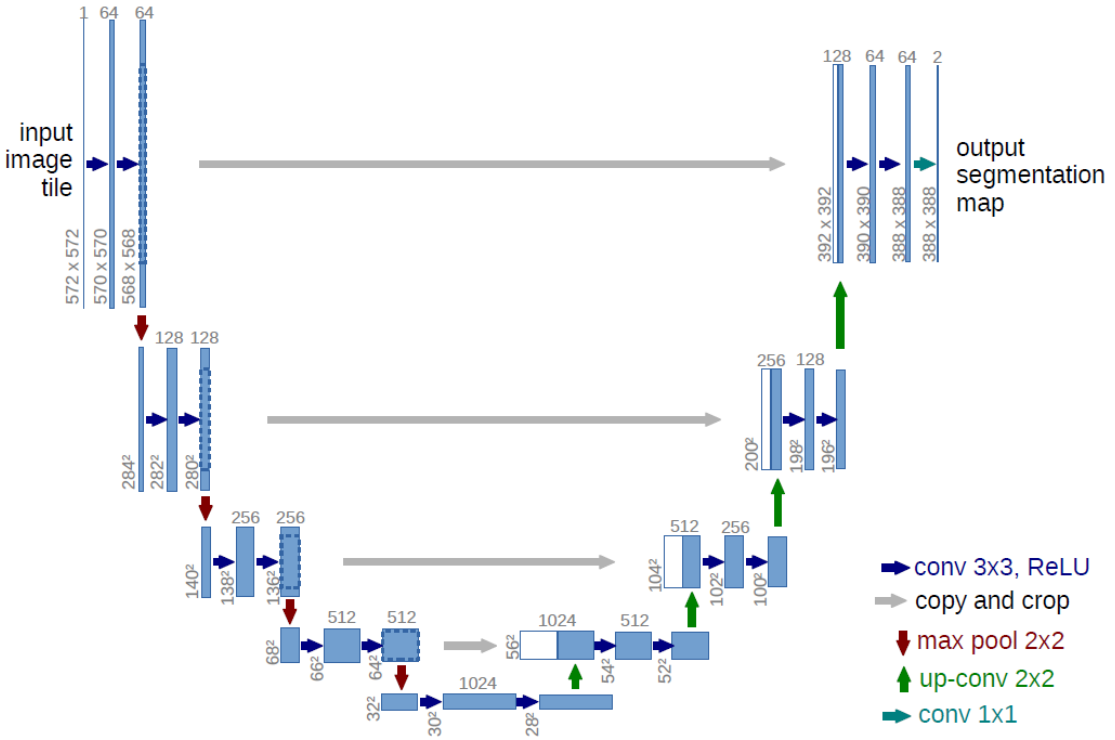

Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

redundancy due to overlapping patches. Secondly, there is a trade-off between localization accuracy and the use of context. Larger patches require more max-pooling layers that reduce the localization accuracy, while small patches allow the network to see only little context. More recent approaches [11,4] proposed a classifier output that takes into account the features from multiple layers. Good localization and the use of context are possible at the same time.

In this paper, we build upon a more elegant architecture, the so-called “fully convolutional network" [9]. We modify and extend this architecture such that it works with very few training images and yields more precise segmentations; see Figure 1. The main idea in [9] is to supplement a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Hence, these layers increase the resolution of the output. In order to localize, high resolution features from the contracting path are combined with the upsampled output. A successive convolution layer can then learn to assemble a more precise output based on this information.

One important modification in our architecture is that in the upsampling part we have also a large number of feature channels, which allow the network to propagate context information to higher resolution layers. As a consequence, the expansive path is more or less symmetric to the contracting path, and yields a u-shaped architecture. The network does not have any fully connected layers and only uses the valid part of each convolution, i.e., the segmentation map only contains the pixels, for which the full context is available in the input image. This strategy allows the seamless segmentation

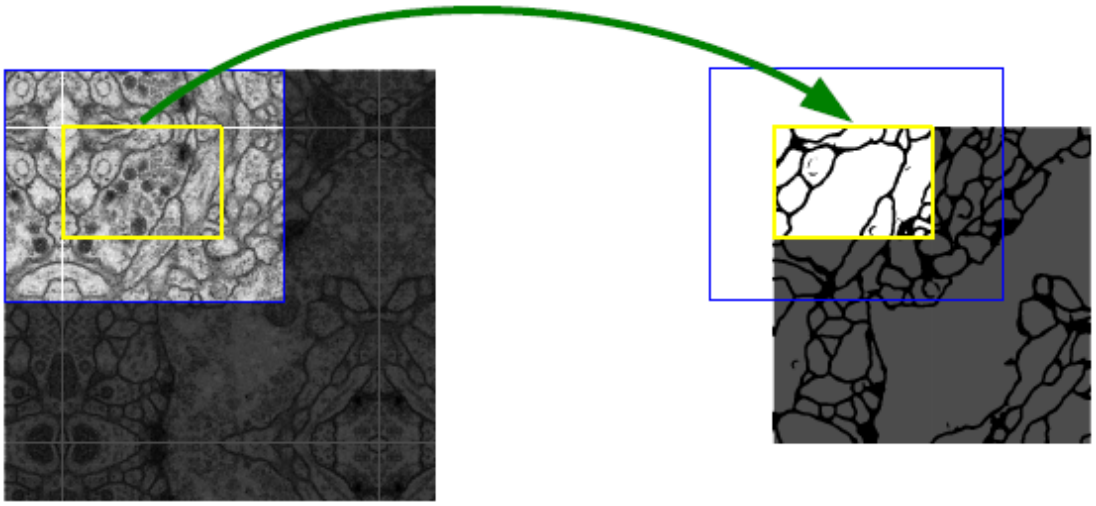

Fig. 2. Overlap-tile strategy for seamless segmentation of arbitrary large images (here segmentation of neuronal structures in EM stacks). Prediction of the segmentation in the yellow area, requires image data within the blue area as input. Missing input data is extrapolated by mirroring

of arbitrarily large images by an overlap-tile strategy (see Figure 2). To predict the pixels in the border region of the image, the missing context is extrapolated by mirroring the input image. This tiling strategy is important to apply the network to large images, since otherwise the resolution would be limited by the GPU memory.

As for our tasks there is very little training data available, we use excessive data augmentation by applying elastic deformations to the available training images. This allows the network to learn invariance to such deformations, without the need to see these transformations in the annotated image corpus. This is particularly important in biomedical segmentation, since deformation used to be the most common variation in tissue and realistic deformations can be simulated efficiently. The value of data augmentation for learning invariance has been shown in Dosovitskiy et al. [2] in the scope of unsupervised feature learning.

Another challenge in many cell segmentation tasks is the separation of touching objects of the same class; see Figure 3. To this end, we propose the use of a weighted loss, where the separating background labels between touching cells obtain a large weight in the loss function.

The resulting network is applicable to various biomedical segmentation problems. In this paper, we show results on the segmentation of neuronal structures in EM stacks (an ongoing competition started at ISBI 2012), where we out-performed the network of Ciresan et al. [1]. Furthermore, we show results for cell segmentation in light microscopy images from the ISBI cell tracking challenge 2015. Here we won with a large margin on the two most challenging 2D transmitted light datasets.

2 Network Architecture

The network architecture is illustrated in Figure 1. It consists of a contracting path (left side) and an expansive path (right side). The contracting path follows the typical

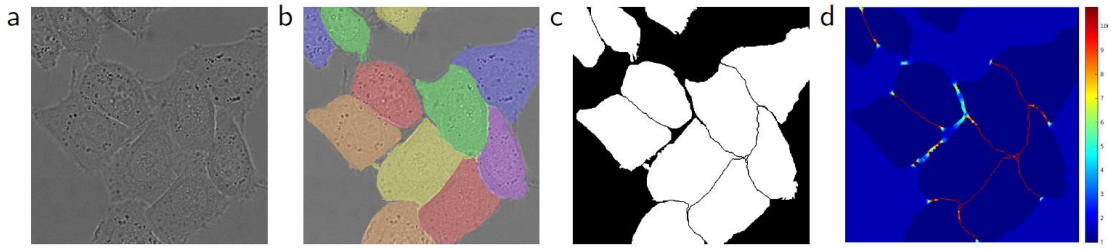

Fig. 3. HeLa cells on glass recorded with DIC (differential interference contrast) microscopy. (a) raw image. (b) overlay with ground truth segmentation. Different colors indicate different instances of the HeLa cells. (c) generated segmentation mask (white: foreground, black: background). (d) map with a pixel-wise loss weight to force the network to learn the border pixels.

architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step we double the number of feature channels. Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution (“up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In total the network has 23 convolutional layers.

To allow a seamless tiling of the output segmentation map (see Figure 2), it is important to select the input tile size such that all 2x2 max-pooling operations are applied to a layer with an even x- and y-size.

3 Training

The input images and their corresponding segmentation maps are used to train the network with the stochastic gradient descent implementation of Caffe [6]. Due to the unpadded convolutions, the output image is smaller than the input by a constant border width. To minimize the overhead and make maximum use of the GPU memory, we favor large input tiles over a large batch size and hence reduce the batch to a single image. Accordingly we use a high momentum (0.99) such that a large number of the previously seen training samples determine the update in the current optimization step.

The energy function is computed by a pixel-wise soft-max over the final feature map combined with the cross entropy loss function. The soft-max is defined as $p_k(x) = \exp(a_k(x)) / (\sum_{k'=1}^{K} \exp(a_{k'}(x)))$ where $a_k(x)$ denotes the activation in feature channel k at the pixel position $x \in \Omega$ with $\Omega \in \mathbb{Z}^2$. K is the number of classes and $p_k(x)$ is the approximated maximum-function. I.e. $p_k(x) \approx 1$ for the k that has the maximum activation $a_k(x)$ and $p_k(x) \approx 0$ for all other k. The cross entropy then penalizes at each position the deviation of $p_{l(x)}(x)$ from 1 using

$$E = \sum_{x \in \Omega} w(x) \log(p_{l(x)}(x)) \tag{1}$$

where $l : \Omega \to \{1, \ldots, K\}$ is the true label of each pixel and $w : \Omega \to \mathbb{R}$ is a weight map that we introduced to give some pixels more importance in the training.

We pre-compute the weight map for each ground truth segmentation to compensate the different frequency of pixels from a certain class in the training data set, and to force the network to learn the small separation borders that we introduce between touching cells (See Figure 3c and d).

The separation border is computed using morphological operations. The weight map is then computed as

$$w(x) = w_c(x) + w_0 \cdot \exp(\frac{-d_1(x) + d_2(x)^2}{2\sigma^2}) \tag{2}$$

where $w_c : \Omega \to \mathbb{R}$ is the weight map to balance the class frequencies, $d_1 : \Omega \to \mathbb{R}$ denotes the distance to the border of the nearest cell and $d_2 : \Omega \to \mathbb{R}$ the distance to the border of the second nearest cell. In our experiments we set $w_0 = 10$ and $\sigma \approx 5 px$.

In deep networks with many convolutional layers and different paths through the network, a good initialization of the weights is extremely important. Otherwise, parts of the network might give excessive activations, while other parts never contribute. Ideally the initial weights should be adapted such that each feature map in the network has approximately unit variance. For a network with our architecture (alternating convolution and ReLU layers) this can be achieved by drawing the initial weights from a Gaussian distribution with a standard deviation of $\sqrt{2/N}$, where N denotes the number of incoming nodes of one neuron [5]. E.g. for a 3x3 convolution and 64 feature channels in the previous layer N = 9*64 = 576.

3.1 Data Augmentation

Data augmentation is essential to teach the network the desired invariance and robustness properties, when only few training samples are available. In case of microscopical images we primarily need shift and rotation invariance as well as robustness to deformations and gray value variations. Especially random elastic deformations of the training samples seem to be the key concept to train a segmentation network with very few annotated images. We generate smooth deformations using random displacement vectors on a coarse 3 by 3 grid. The displacements are sampled from a Gaussian distribution with 10 pixels standard deviation. Per-pixel displacements are then computed using bicubic interpolation. Drop-out layers at the end of the contracting path perform further implicit data augmentation.

4 Experiments

We demonstrate the application of the u-net to three different segmentation tasks. The first task is the segmentation of neuronal structures in electron microscopic recordings. An example of the data set and our obtained segmentation is displayed in Figure 2. We provide the full result as Supplementary Material. The data set is provided

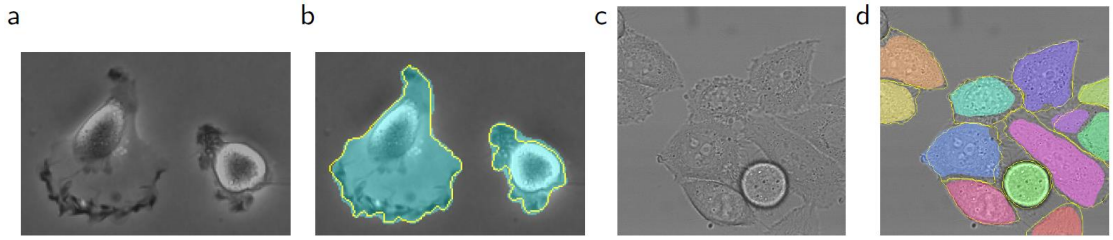

Fig. 4. Result on the ISBI cell tracking challenge. (a) part of an input image of the "PhC-U373"data set. (b) Segmentation result (cyan mask) with manual ground truth (yellow border) (c) input image of the "DIC-HeLa" data set. (d) Segmentation result (random colored masks) with manual ground truth (yellow border).

by the EM segmentation challenge [14] that was started at ISBI 2012 and is still open for new contributions. The training data is a set of 30 images (512x512 pixels) from serial section transmission electron microscopy of the Drosophila first instar larva ventral nerve cord (VNC). Each image comes with a corresponding fully annotated ground truth segmentation map for cells (white) and membranes (black). The test set is publicly available, but its segmentation maps are kept secret. An evaluation can be obtained by sending the predicted membrane probability map to the organizers. The evaluation is done by thresholding the map at 10 different levels and computation of the "warping error", the "Rand error" and the "pixel error"[14].

The u-net (averaged over 7 rotated versions of the input data) achieves without any further pre- or postprocessing a warping error of 0.0003529 (the new best score, see Table 1) and a rand-error of 0.0382.

This is significantly better than the sliding-window convolutional network result by Ciresan et al. [1], whose best submission had a warping error of 0.000420 and a rand error of 0.0504. In terms of rand error the only better performing algorithms on this data set use highly data set specific post-processing methods1 applied to the probability map of Ciresan et al. [1].

Table 1. Ranking on the EM segmentation challenge [14] (march 6th, 2015), sorted by warping error.

Rank	Group name	Warping Error	Rand Error	Pixel Error
	** human values **	0.000005	0.0021	0.0010
1.	u-net	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA [1]	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
⋮				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

We also applied the u-net to a cell segmentation task in light microscopic images. This segmentation task is part of the ISBI cell tracking challenge 2014 and 2015 [10,13]. The first data set "PhC-U373" contains Glioblastoma-astrocytoma U373 cells on a polyacrylimide substrate recorded by phase contrast microscopy (see Figure 4a,b and

Table 2. Segmentation results (IOU) on the ISBI cell tracking challenge 2015.

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

Supp. Material). It contains 35 partially annotated training images. Here we achieve an average IOU ("intersection over union") of 92%, which is significantly better than the second best algorithm with 83% (see Table 2). The second data set "DIC-HeLa"3 are HeLa cells on a at glass recorded by differential interference contrast (DIC) microscopy (see Figure 3, Figure 4c,d and Supp. Material). It contains 20 partially annotated training images. Here we achieve an average IOU of 77.5% which is significantly better than the second best algorithm with 46%.

5 Conclusion

The u-net architecture achieves very good performance on very different biomedical segmentation applications. Thanks to data augmentation with elastic deformations, it only needs very few annotated images and has a very reasonable training time of only 10 hours on a NVidia Titan GPU (6 GB). We provide the full Caffe[6]-based implementation and the trained networks4. We are sure that the u-net architecture can be applied easily to many more tasks.

Acknowledgements

This study was supported by the Excellence Initiative of the German Federal and State governments (EXC 294) and by the BMBF (Fkz 0316185B).

References

1. Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS. pp. 2852–2860 (2012)
2. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
4. Hariharan, B., Arbelez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization (2014), arXiv:1411.5752 [cs.CV]
5. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification (2015), arXiv:1502.01852 [cs.CV]
6. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding (2014), arXiv:1408.5093 [cs.CV]
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)
8. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation 1(4), 541–551 (1989)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation (2014), arXiv:1411.4038 [cs.CV]
10. Maska, M., (...), de Solorzano, C.O.: A benchmark for comparison of cell tracking algorithms. Bioinformatics 30, 1609–1617 (2014)
11. Seyedhosseini, M., Sajjadi, M., Tasdizen, T.: Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 2168–2175 (2013)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014), arXiv:1409.1556 [cs.CV]
13. WWW: Web page of the cell tracking challenge, http://www.codesolorzano.com/celltrackingchallenge/Cell_Tracking_Challenge/Welcome.html
14. WWW: Web page of the em segmentation challenge, http://brainiac2.mit.edu/isbi_challenge/

附录 II 文献翻译译文
U-Net：生物医学图像分割的卷积网络

Olaf Ronneberger，Philipp Fischer 和 Thomas Brox
计算机科学系和 BIOSS 生物信号研究中心，
德国弗赖堡大学
ronneber@informatik.uni-freiburg.de，
主页：http://lmb.informatik.uni-freiburg.de/

摘要。 通常认为，深层网络的成功训练需要许多带标记的训练样本。在这篇论文中，我们提出了一个新的网络和训练策略。为了更有效的利用标注数据，我们使用数据增强的方法(data augmentation)。该网络由两部分组成：一个收缩路径(contracting path)来获取上下文信息以及一个对称的扩张路径(expanding path)用以精确定位。我们展示了这样的网络可以用比较少的图片进行端到端的训练，我们使用这个网络获得了赢得了 ISBI cell tracking challenge 2015。不仅如此，这个网络非常的快，对一个 512*512 的图像，使用一块现代的 GPU 只需要不到一秒的时间。

1 介绍

在过去的几年中，深度卷积网络在许多识别任务上获得了当前最好的结果。卷积网络已经存在很多年了，但因为训练集的大小和网络的大小的有限，它们的成就被限制住了。后来获得的突破是，Krizhevsky 等人在一个含有几百万个参数的 8 层网络上，通过包含 100 万张训练图像的 ImageNet 数据集来进行监督训练。在这之后，更大更深的网络被用来训练。

卷积神经网络的典型用处是用在分类任务上，即一张图像的输出结果是一个类别标签。但是，在很多视觉的任务上，尤其是生物图像的处理上，想得到的结果应该是包含定位的，也就是说，对每个像素都应该做一个分类。而且，几千张训练图像在生物领域中通常是很难得到的。因此，Ciresan 等人通过提供该像素的局部区域作为输入，来训练了一个带滑动窗口的网络，预测每个像素的类别标签。首先，这个网络可以用来定位，其次，就 patch 数而言训练数据的量远大于训练图像的量。这个网络大幅度赢得了 EM segmentation challenge at ISBI 2012。

这个网络有两个很明显的缺点：1.它很慢，因为要分别预测每一个 patch 的类别，patch 之间的重叠导致每次预测都要重复计算同一个点。2.这个网络需要在局部准确性和获取整体上下文信息之间取舍。大的 patches 需要更多的 max-pooling 层，导致定位精度下降，但小的 patches 使网络只能获得很少的上下文信息。最近提出的方法把多层的特征考虑进去，做一个分类输出。好的定位和利用上下文信息可以同时做到。

在这篇文章中，我们建立了一个更加优雅的框架，通常被称为“全卷积网络”(fully convolutional network)。我们修改并拓展了这个框架，使其可以仅使用少量训练图片就可以工作，获得更高的分割准确率。网络图 1 所示。[9]中的主要

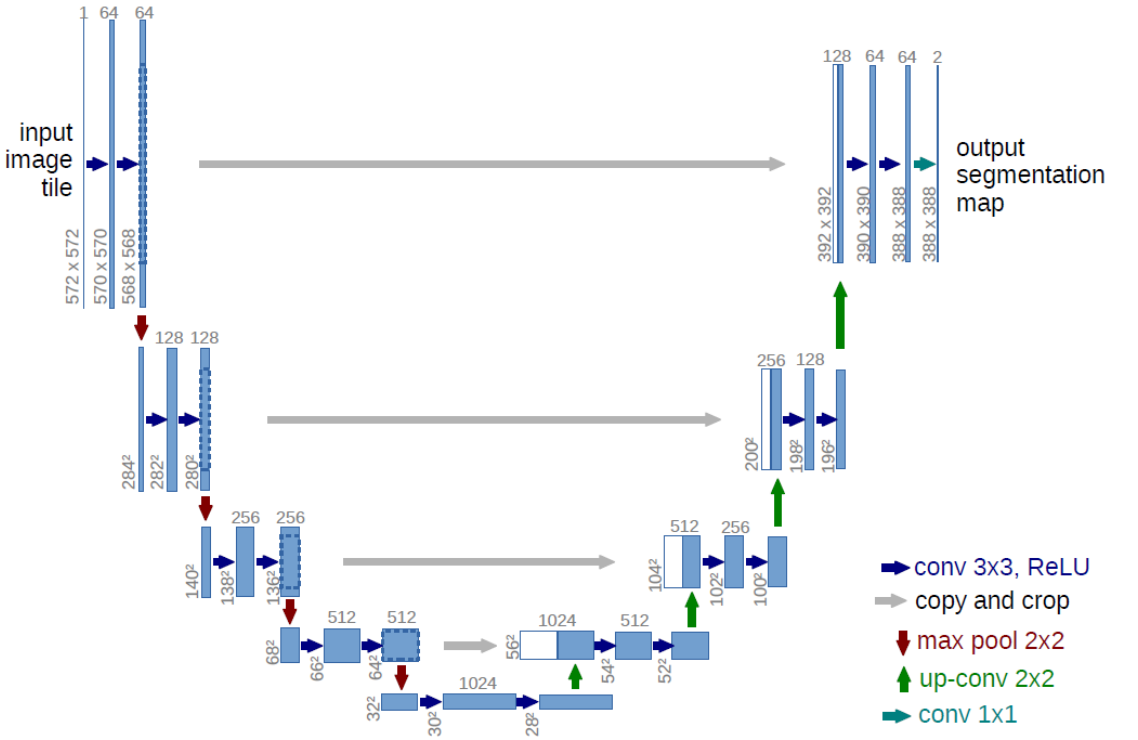

图 1：U-net 架构(以 32*32 的最低分辨率为例)。每个蓝色框对应一个多通道特征图。通道数量在框的顶部。x-y 大小位于框的左下角。白色框表示复制后的特征图。箭头表示不同的操作。

思想是通过补充一个常用的收缩网络连续层，pooling 层被上采样层所取代。因此，这些层增加了输出的分辨率。为了定位，来自收缩路径的高分辨率特征与上采样输出相结合。一个连续的卷积层可以根据这些信息学习组装一个更精确输出。

在我们的架构中一个重要的修改是在上采样中，我们有大量的特征通道数量，这些通道允许网络将上下文信息传播到更高分辨率的层。所以，扩张路径与收缩路径或多或少是对称的，并且产生一个 U 形的结构。该网络没有任何全连接层，并且仅使用每个卷积的有效部分，即仅使用分割图，这个分割图包含输入图像中完整上下文可用的像素。这种策略通过 overlap-tile 策略(图 2)，可以允许无缝分割任意大的图像。为了预测边界区域中的像素，我们通过镜像输入图像来推断丢失的上下文。这种平铺策略对于将网络应用于大尺寸图像很重要，因为不这样做的话，图像的分辨率将受到 GPU 内存的限制。

至于我们的任务，只有很少的训练数据可用，我们将使用很多的数据增强，也就是将弹性变形施加到训练图像。这样可以让网络学习到对这种变形的不变性，而不需要通过在标记图像语料库中看到这些转换。这在生物医学分割中尤其重要，因为变形是组织中最常见的变化，实际变形可以被高效的模拟。Dosovitskiy 等人在无监督特征学习中，将学习不变性的数据增强的价值展示了出来。

在细胞分割任务中的另一个挑战是，如何将同类别的相互接触的目标分开(图 3)。我们提出了使用一种带权重的损失(weighted loss)。在损失函数中，分割相互接触的细胞获得了更大的权重。

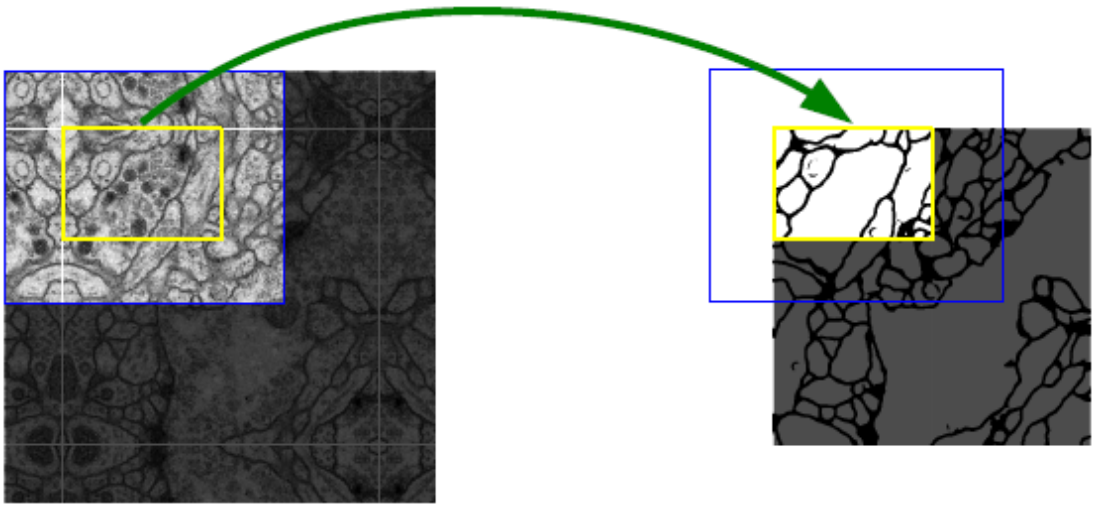

图 2：无缝分割任意大图像的重叠拼贴策略（此处为 EM 堆栈中的神经元结构的分割）。预测黄色区域的分割需要蓝色区域内的图像数据作为输入。缺少输入数据是通过镜像来推断的。

最终的网络适用于各种生物医学分割问题。在本文中，我们展示了在 EM 套件中关于神经结构分割的结果(ISBI 2012 开始的持续竞争)中，我们跑赢了 Ciresan 等人的网络。此外，我们展示了的来自 ISBI 细胞的光学显微镜图像的细胞分割结果。利用这个结果，我们大幅度赢得了两场最具挑战性的 2D 透射光数据集赛事。

2 网络结构

图 1 展示了网络结构，它由收缩路径和扩张路径组成。收缩路径是典型的卷积网络架
构。它的架构是一种重复结构，每次重复中都有 2 个卷积层和一个 pooling 层，卷积层中卷积核大小均为 3*3，激活函数使用 ReLU，两个卷积层之后是一个 2*2 的步长为 2 的 max pooling 层。每一次下采样后我们都把特征通道的数量加倍。收缩路径中的每一步都首先使用反卷积(up-convolution)，每次使用反卷积都将特征通道数量减半，特征图大小加倍。反卷积过后，将反卷积的结果与收缩路径中对应步骤的特征图拼接起来。收缩路径中的特征图尺寸稍大，将其修剪过后进行拼接。对拼接后的 map 进行 2 次 3*3 的卷积。最后一层的卷积核大小为 1*1，将 64 通道的特征图转化为特定深度(分类数量，二分类为 2)的结果。网络总共 23 层。

为了实现分割图的无缝平铺（参见图 2），选择输入切片大小非常重要，以便将所有 2x2 的 max-pooling 操作应用于具有偶数 x 和 y 大小的图层。

3 训练

输入图像及其对应的分割图被用于训练。该网络随着随机梯度下降实现。由于不带边界的卷积，输出图像比输入小一个恒定的边界宽度。为尽量减少开销并尽量利用的 GPU 内存，我们倾向于在大批量的情况下使用大量输入 tiles，因此，将

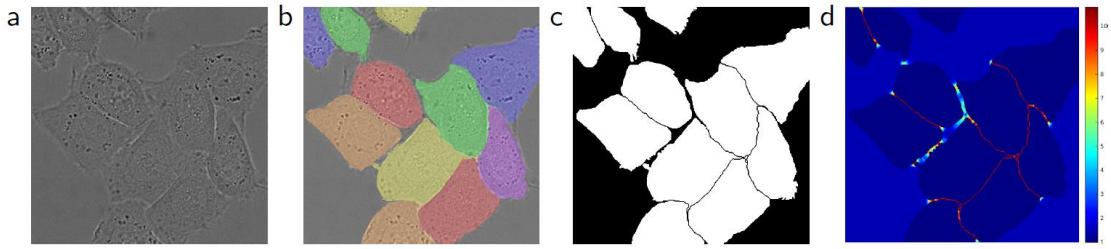

图 3：用 DIC 光学显微镜记录的玻璃上的 HeLa 细胞。(a)原始图像。(b)将 ground truth 分割叠加到原始图像。(c)生成的分割图(白色：前景；黑色：背景)。(d)像素级的损失权重图，用来强制使网络学到边界。

批次减少为单个图像。因此我们使用高动量(0.99)，这样利用了大量以前的训练样本来确定在当前优化步骤中的参数更新。

能量函数通过在最后一层特征图上的像素级别的 soft-max 结合交叉熵损失函数来计算。Soft-max 定义为 $p_k(x)=\exp(a_k(x))/(\sum_{k'=1}^{K}\exp(a_{k'}(x)))$，其中 $a_k(x)$ 为激活函数，在第 k 个特征图通道，像素位置为 x 处；K 为分类的个数；$p_k(x)\approx 1$ 表示该通道的特征图为该类别的概率是最大的。交叉熵对每个位置的 $p_{l(x)}(x)$ 距离 1 的偏差做一个惩罚：

$$E=\sum_{x\in\Omega} w(x)\log(p_{l(x)}(x))$$

其中 $l:\Omega\to\{1,\ldots,K\}$ 为每个像素的真实标签，$w:\Omega\to\mathbb{R}$ 是一个权重图，利用这个权重图，在训练时可以给某些像素更高的权重。

我们事先计算每个 ground truth 分割的权重图来补偿训练集中某一特定种类中的像素的不同出现频率，同时也是为了强制使网络学到互相接触的细胞之间的细小的边界(图 3)。

分割边界用形态学操作来计算，权重图的计算如下：

$$w(x)=w_c(x)+w_0\cdot\exp\left(\frac{-d_1(x)+d_2(x)^2}{2\sigma^2}\right)$$

其中 $w_c:\Omega\to\mathbb{R}$ 用来平衡类之间频率的权重图，$d_1:\Omega\to\mathbb{R}$ 表示到最近细胞边界的距离，$d_2:\Omega\to\mathbb{R}$ 表示离第二近的。在我们的操作中，我们设定 $w_0=10$ ，$\sigma\approx 5px$ 。

在带有许多卷积层和不同路径的深度网络中，一个好的初始化权重是非常重要的，否则网络的一部分可能过于激活，而另一部分没有做贡献。理想的初始化权重应当满足不同的特征图之间有单位的不相同。在我们的网络(选择性卷积和 ReLU 层)中，可以通过从高斯分布来满足。标准差为 $\sqrt{2/N}$ ，其中 N 为该神经单元的输入节点的个数。对于一个 64 通道的特征图，3×3 的卷积核，$N=9\times 64=576$。

3.1 数据增强

当只有少数训练样本可用时，数据增强对教给网络学到所需的不变性和鲁棒性至关重要。对于显微镜图像我们主要需要对偏移和旋转的不变性以及对变形和灰

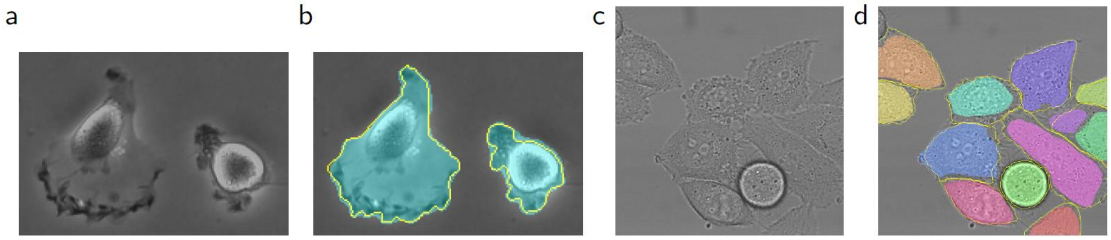

图 4：ISBI 细胞跟踪挑战的结果。(a)“PhC-U373”数据集的输入图像的一部分。(b)带人工标记 ground truth(黄色边框)的分割结果(青色遮罩)(c)“DIC-HeLa”数据集的输入图像。 (d)带人工标记 ground truth(黄色边框)的分割结果(随机彩色遮罩)。

度值变化的鲁棒性。 特别是训练样本的随机弹性变形，它似乎是利用很少的带标记图像来训练一个分割网络的关键。我们使用随机位移矢量在粗糙的 3×3 网格上生成平滑变形。位移是从 10 像素标准差的高斯分布中采样得到。然后使用双三次插值计算每像素位移。收缩路径末端的 drop-out 层来执行隐式表达的数据增强。

4 实验

我们展示了 u-net 在三种不同分割中的应用任务。第一项任务是电子显微镜记录的图像中神经元结构的分割。数据集和我们获得的分割的一个例子显示在图 2 中。我们提供完
整的结果作为补充材料。该数据集由 EMB 细分挑战提供，该项挑战始于 2012 年 ISBI，并且对新的贡献者仍然开放。训练数据是来自果蝇第一龄幼虫腹侧神经索(VNC)的连续切片透射电子显微镜的一组 30 个图像(512x512 像素)。每个图像都带有相应的完全标记的细胞(白色)和膜(黑色)的 ground truth 分割图。测试集是公开可用的，但其分割图是未知的。通过将预测的膜概率图发送给组织者可以获得评估结果。评估是通过 10 个不同级别对分割图进行阈值化并计算“翘曲误差”，“随机误差”和“像素误差”。

U-net(对输入数据的 7 个旋转版本进行平均)在没有进一步预处理或后处理的情况下达到 0.0003529 的翘曲误差(新的最佳得分，参见表 1)和 0.0382 的随机误差。

这比 Ciresan 等人的滑动窗口卷积网络结果要好得多，其最佳提交的翘曲误差为 0.000420，随机误差为 0.0504。就随机误差而言，该数据集上唯一性能更好的算法使用高度数据集特定的后处理方法 1 应用于 Ciresan 等人的概率图。

我们还将 u-net 应用于光学显微图像中的细胞分割任务。该分割任务是 ISBI 单元跟踪挑战 2014 和 2015 的一部分。第一组数据集“PhC-U373”包含通过相差显微术记录在聚丙烯酰亚胺基质上的成胶质细胞瘤-星形细胞瘤 U373 细胞(参见图 4a，b 和补充材料)。它包含 35 个部分标记的训练图像。在这里，我们实现了 92％的平均 IOU(“交并比”)，这比 83％的次优算法要好得多(见表 2)。第二组数据集“DIC-HeLa”是通过差分干涉对比(DIC)显微镜记录的玻璃上的 HeLa 细胞(参见图 3，图 4c，d 和补充材料)。它包含 20 个部分标记的训练图像。 这里我们实现了 77.5％的平均 IOU，明显好于 46％的第二优秀算法。

表 1：EM 细分挑战排名[14](2015 年 3 月 6 日)，按翘曲误差排序。

Rank	Group name	Warping Error	Rand Error	Pixel Error
	** human values **	0.000005	0.0021	0.0010
1.	u-net	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA [1]	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
⋮				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

表 2：ISBI 细胞追踪挑战 2015 的分割结果(IOU)。

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

5 结论

U-net 架构在不同的生物医学分割应用中取得了非常好的性能。由于具有弹性变形的数据增强，它只需要很少的标记图像，并且在 NVidia Titan GPU(6 GB)上的训练时间非常短，只需要 10 个小时。我们提供了完整的基于 Caffe 的实现和训练好的网络。我们相信，u-net 架构可以轻松应用于更多的任务。

致谢

这项研究得到了德国联邦和州政府卓越计划(EXC 294)和 BMBF(Fkz 0316185B)的支持。

引用

1. Ciresan, D.C., Gambardella, L.M., Giusti, A., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS. pp. 2852–2860 (2012)
2. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
4. Hariharan, B., Arbelez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization (2014), arXiv:1411.5752 [cs.CV]
5. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification (2015), arXiv:1502.01852 [cs.CV]
6. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding (2014), arXiv:1408.5093 [cs.CV]
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)
8. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation 1(4), 541–551 (1989)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation (2014), arXiv:1411.4038 [cs.CV]
10. Maska, M., (...), de Solorzano, C.O.: A benchmark for comparison of cell tracking algorithms. Bioinformatics 30, 1609–1617 (2014)
11. Seyedhosseini, M., Sajjadi, M., Tasdizen, T.: Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 2168–2175 (2013)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014), arXiv:1409.1556 [cs.CV]
13. WWW: Web page of the cell tracking challenge, http://www.codesolorzano.com/celltrackingchallenge/Cell_Tracking_Challenge/Welcome.html
14. WWW: Web page of the em segmentation challenge, http://brainiac2.mit.edu/isbi_challenge/