Journal of System Simulation

Volume 32 | Issue 7 Article 12

7-15-2020

Human Depth Maps Restoration Based on Guided GAN

Jingfang Yin
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;;

Dengming Zhu
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;;3. Taicang-CAS Institute of Information and Technology, Taicang 215400, China;

Shi Min
2. University of Chinese Academy of Sciences, Beijing 102206, China;;

Zhaoqi Wang
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;;3. Taicang-CAS Institute of Information and Technology, Taicang 215400, China;

Human Depth Maps Restoration Based on Guided GAN

Abstract

Abstract: The depth maps captured by a small depth camera on mobile devices suffer from the problem of severe holes. The *Guided Generative Adversarial Network (Guided GAN)* based on deep learning is proposed to restore human depth maps with above problems. The high-precision human segmentation features and depth class features are extracted from the monocular RGB image by the guider based on the stacked hourglass network. *The holes in the human depth maps are filled by the special generator under the guidance of the extracted human features. In order to get the more realistic results, the discriminator is introduced* to optimize the generator. The experimental results show that the proposed method can restore the human depth maps effectively in the existing human datasets and the dataset collected by the small depth camera. It achieves better results than the existing method.

Keywords

RGBD camera, human depth data restoration, deep learning, two-stage stacked hourglass network, guided GAN

Recommended Citation

Yin Jingfang, Zhu Dengming, Shi Min, Wang Zhaoqi. Human Depth Maps Restoration Based on Guided GAN[J]. Journal of System Simulation, 2020, 32(7): 1312-1321.

第 32 卷第 7 期 系统仿真学报© Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

基于引导对抗网络的人体深度图像修补方法

阴敬方[1]，朱登明[1,3,*]，石敏[2]，王兆其[1,3]
(1. 中国科学院计算技术研究所，北京 100190；2. 华北电力大学，北京 102206；
3. 太仓中科信息技术研究院，江苏 太仓 215400)

摘要：移动设备配备的小型深度相机采集到的人体深度图像存在严重的孔洞问题。针对该问题，提出基于深度学习的*引导对抗网络*。使用基于堆叠沙漏网络的引导器从 RGB 图像中提取人体部分分割特征和深度类别特征；*在上述人体特征引导下，使用独特的生成器修复人体深度图像中的孔洞。为使结果更加逼真，加入判别器在网络训练过程中对生成器进行优化调整。*实验结果显示，该方法在现有的人体数据集和小型深度相机采集的数据集上，都能很好解决孔洞问题，均取得比现有方法更好的效果。

关键词：深度相机；人体深度图像修复；深度学习；堆叠沙漏网络；引导对抗网络
中图分类号：TP391.9 文献标识码：A 文章编号：1004-731X (2020) 07-1312-10
DOI: 10.16182/j.issn1004731x.joss.19-VR0462

Human Depth Maps Restoration Based on Guided GAN

Yin Jingfang[1], *Zhu Dengming*[1,3*], *Shi Min*[2], *Wang Zhaoqi*[1,3]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 102206, China; 3. Taicang-CAS Institute of Information and Technology, Taicang 215400, China)

Abstract: The depth maps captured by a small depth camera on mobile devices suffer from the problem of severe holes. The *Guided Generative Adversarial Network (Guided GAN)* based on deep learning is proposed to restore human depth maps with above problems. The high-precision human segmentation features and depth class features are extracted from the monocular RGB image by the guider based on the stacked hourglass network. *The holes in the human depth maps are filled by the special generator under the guidance of the extracted human features*. *In order to get the more realistic results, the discriminator is introduced* to optimize the generator. The experimental results show that the proposed method can restore the human depth maps effectively in the existing human datasets and the dataset collected by the small depth camera. It achieves better results than the existing method.

Keywords: RGBD camera; human depth data restoration; deep learning; two-stage stacked hourglass network; guided GAN

引言

人体深度是深度信息中极为重要的一种，它是指视点到人体的距离，通常用于人体骨架提取、三维人体重建和非接触式人体测量。在虚拟试衣、人机交互、游戏、影视等诸多领域，人体深度有着广泛的应用。

目前，人体深度信息普遍采用深度相机来获取，但深度相机获取人体深度仍存在一些问题，例如：受测量范围的限制和环境光的影响，深度相机

收稿日期：2019-08-29 修回日期：2019-11-13；
基金项目：国家自然科学基金(61532002)，国家科技重大专项(2017ZX05019005)，太仓大院大所项目(TC2017DYDS07)；
作者简介：阴敬方(1997-)，男，山东莱州，硕士生，研究方向为虚拟现实、计算机图像与图形；朱登明(通讯作者 1973-)，男，安徽太湖，博士，副研究员，研究方向为虚拟现实、计算机图形学。

采集到的深度图中会存在大量孔洞；深度图的采集和传输过程中，会因干扰而产生噪声。这些问题在移动设备配备的小型深度相机中尤其明显。严重的孔洞问题和噪声问题使得深度图的质量下降，直接影响以它为源数据的人体建模和非接触式人体测量等应用。因此，在深度相机小型化、移动化的发展趋势下，获取高质量的人体深度图成为计算机视觉领域重要的研究方向之一。

人体深度图中的噪声大多是干扰导致的椒盐噪声，简单的中值滤波方法[1]可以将其修复。不同于噪声问题，人体深度图中的孔洞问题仍未得到很好的解决。

孔洞修复问题早在照片修复领域得到了研究者的关注，并诞生了一系列方法，例如：基于扩散的 (diffusion-based)[2] 、基于块匹配的 (Patch match)[3]、基于图像融合的(image melding)[4]修复方法。这些方法都取得了不错的效果。最近由于机器学习的发展和深度学习的广泛运用，基于卷积神经网络的 (CNN-based)[5-7] 和基于对抗神经网络 (GAN-based)[8]的照片补全方法也取得了显著成果。但这些方法都是针对三通道彩色(RGB)图像设计，而 RGB 图像与深度相机采集到的深度图像在数值范围、数值表示意义、数值连续性等方面有很大的区别。这导致了 RGB 图像的修补方案难以泛化到修补人体深度图。

为解决人体深度图中的孔洞问题，本文提出基于引导对抗网络的方法对其进行修补。不同于普通的生成对抗网络 (Generative Adversarial Networks)[9]，引导对抗网络受人体特征的引导来修复人体深度图。该网络首先从人体数据集中学习到人体部分分割、相对人体深度特征，之后在这些特征的引导下修复人体深度图。其结构包含三个子网络：引导器、生成器、判别器。其中，引导器基于堆叠沙漏网络(Stacked Hourglass network)[10]而设计，以深度相机拍摄的彩色图像为输入来预测人体部分分割和相对人体深度类别，其预测结果为后续网络提供人体特征上的指导。在堆叠沙漏网络的内部，本文使用反卷积代替最邻近元插值的上采样方法。引导对抗网络的生成器用来修复人体深度图。不同于普通对抗网络的生成器，其输入包含 2 部分：引导器的预测结果和深度相机拍摄的深度图，并且以不同的网络层对两输入分别进行特征提取。判别器用来鉴定生成器输出结果的真伪，从而指导生成器产生更逼真的人体深度图。

为使得模型更好地理解人体并具有更高的泛化能力，本文对训练集数据进行了数据增强。除此之外，受迁移学习的启发，模型训练采用预训练与微调相结合的方式，以防止过拟合。模型训练后，在人体数据集和深度相机采集的数据集上进行了测试，并对结果进行了量化分析。

1 相关工作

1.1 人体理解

基于 RGB 图像的人体内容理解，包含的典型研究方向有：包围盒(Bounding Box)[11]，二维姿势估计(2D Pose Estimation)[10,12-15]，三维姿势估计(3D Pose Estimation)[16-17]，人体分割(Segmentation)[18-20]和人体深度预测(Depth Prediction)[20]。

针对 RGB 图像中的人体分割，传统方法大多采用人体轮廓和人体二维姿势之间的几何关系进行求解[21-22]。近年来，随着深度学习的兴起，很多研究者设计了用于人体分割的网络结构，取得了很好的效果。Long 等[18]提出的全卷积网络(Fully Convolutional Network)可用于人体语义分割。Newell 等提出了堆叠沙漏网络(Stacked hourglass network)[10]最初用来进行人体二维姿态预测，并在 FLIC 和 MPII 两个数据集上取得了当时最好的成绩。Varol 等[20]基于堆叠沙漏网络，实现了人体部件分割，取得不错效果。Zhao 等[19]提出的人体解析数据集融合了包围盒目标检测、人体语义分割，可用于人体服饰语义分析。本文网络在人体部分分割模块以 Varol 等的工作为基础，受全卷积网络[17]的启发，使用反卷积改进堆叠沙漏网络。本文网络

DOI: 10.16182/j.issn1004731x.joss.19-VR0462

第 32 卷第 7 期 系统仿真学报 Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

整体受人体服饰语义分割[19]的启发，在保证整体端到端的同时，局部分阶段设计。

大规模人体数据集的出现为深度学习方法研究人体提供了可行性。Ionescu 等[23]提出的室内人体数据集 Human3.6M 包含人体分割、人体深度、二维姿态、三维姿态的等信息。该数据集由 11 位演员，在室内模仿 15 种实现情景采集而成。Varol 等[20]提出大规模虚拟人体数据集 SURREAL。该数据集是用 SMPL[24]人体模型，在多种形状、纹理、视角和姿态上的变化下进行渲染而成。本文将上述两个人体数据集经过针对性的预处理和增强后，用于网络模型训练。

1.2 基于 RGB 图像的深度预测及优化

针对 RGB 图像中人体的深度预测，之前的研究[25-27]大都面向一般场景进行深度预测，Varol 等[20]基于沙漏网络(Hourglass network)，实现了 64×64 分辨率的相对人体深度类别预测。本文方法在人体深度类别预测部分，改进了 Varol 等的方法，使得模型对人体特征的理解更充分，以提高后续的人体深度图的修补质量。

Liu 等[28]提出了一个基于 RGB 图像的深度图恢复优化框架，解决了彩色图与深度图边缘不一致问题和深度图中的深度值不连续的问题。其工作成果证明了 RGB 图在深度修复中的指导作用，为本文整体思路提供了参考。

1.3 图像修复

孔洞和缺失问题是照片修复领域的关键性问题，大量照片修复方法为解决此问题而提出。传统的方法主要有基于扩散的(diffusion-based)方法[2]、基于块匹配(Patch match)的方法[3]、基于图像融合的(image melding)[4]、基于大型图片数据库的方法[29]等等。传统方法都有较为明显的局限性，例如，基于扩散的方法一般只能填补小洞或窄孔，基于小块的方法生成的纹理都来自于输入的图像，基于大型图片数据库的方法需要假定数据库中包含与输入图像类似的图像。

随着深度学习的发展，卷积神经网络(CNN)也被应用于图像修补领域。近年来比较经典的是基于生成对抗网络的上下文编码器(Context Encoder)[8]，该方法使用全卷积 L2 损失(L2 Loss)与对抗损失(GAN Loss)结合来训练网络，在 128×128 像素大小的图像中心修复 64×64 像素大小的区域。该方法是特征学习驱动的，在修复较大的图像缺失的同时保持恢复好的图像符合整幅图像的语义。但是该方法修复后的图像会存在局部模糊问题，导致肉眼容易辨别生成的图片与真实的图片。Satoshi 等[30]提出了一种新颖的图片修复方法，此方法以全卷积网络为基础，结合全局和局部对抗损失，完成图像修补。这种方法可以填充任何形状的缺失区域来补全任意分辨率的图像，同时保持图像在局部和全局一致，但是这种方法依然针对 RGB 图像而设计，无法很好的解决人体深度修复的问题。受这些方法的启发，本文将生成对抗网络的生成器和判别器模块加以改进，很好地运用到引导对抗网络中。

2 方法提出

为解决小型深度相机拍摄人体的深度图中存在的孔洞问题，本文提出引导对抗网络，其整体架构如图 1 所示。该网络可分为 2 大阶段：人体深度预测、人体深度修复，其中人体深度预测阶段的结果作为后续网络输入的一部分，为后续网络提供引导。人体深度预测阶段的任务是从 RGB 图像中预测人体分割和相对人体深度类别，由引导器完成。人体深度修复阶段的任务是在深度类别的指导下修复人体深度图并使得结果符合人体特性，由生成器和判别器完成。

2.1 人体深度类别预测

人体深度类别是预测根据人体先验知识，将人体深度分成的多个区间，即将连续的人体深度值转化为离散的人体深度类别。

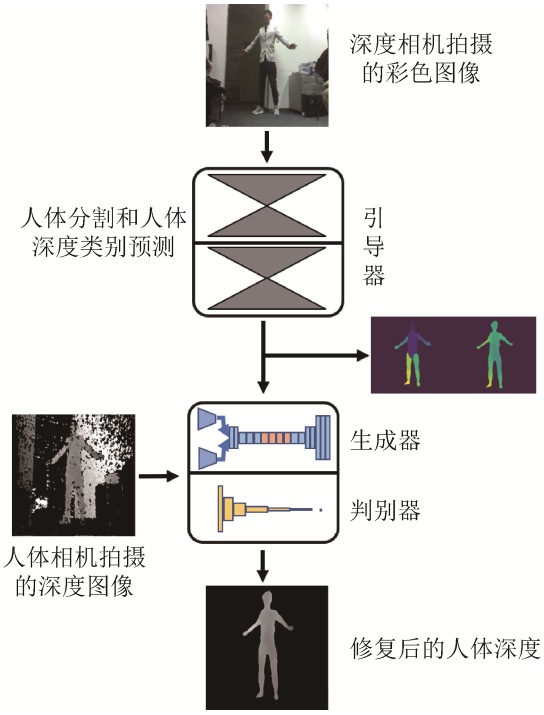

图 1　引导对抗网络整体架构图
Fig. 1　Overall architecture of Guided GAN

人体类别预测部分实质是像素级分类问题，由引导器完成。由于人体是一种局部刚性物体，具有相对不变的局部特征和多变的全局特征，直接预测人体深度类别是比较困难的，所以在已知人体分割的前提下预测人体深度类别会使得计算相对容易。相比于 Varol[20]的人体深度预测工作，引导器将人体部分分割模块和人体深度预测模块串接成一个整体的网络，以人体部分分割结果为指导实现人体深度类别预测。除此之外，引导器使用反卷积层代替在沙漏网络中传统的最邻近差值，实现更精准的上采样。

引导器网络结构如图 2 所示，由卷积层、堆叠沙漏层、反卷积输出层构成，其中每个堆叠沙漏层都由两个沙漏网络堆叠组成。输入为一个 3 通道 256×256 大小的 RGB 图像，通过卷积层对其提取特征，将其变成 64×64 特征图，通过沙漏网络，可得到人体部分分割和深度类别预测的结果，其大小分别为 15×64×64 和 20×64×64。在网络的最后，增加反卷积操作，将输出的分辨率提升至 15×256×256 和 20×256×256。人体部分分割结果的通道数为 15，包括 14 类人体部分和背景；人体深度预测结果的通道数为 20，包含 19 类人体深度和背景。

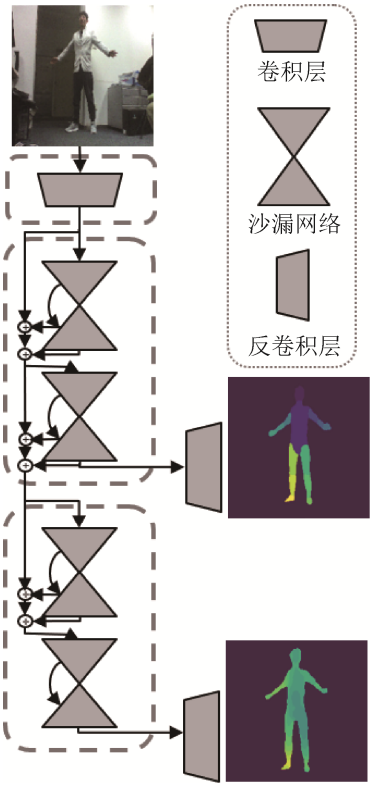

图 2　引导器网络整体架构图
Fig. 2　Architecture of guider

堆叠沙漏层采用了四阶沙漏来实现多尺度特征的提取。Chen 等[31]将沙漏网络用于人脸超分辨任务，并且证明了，沙漏数量增加时，先验信息估计网络结构越深，学习能力越强，性能越好。在人体深度类别预测任务中，不同数量的沙漏网络单元的实验结果比较如表 1 所示，表中所用的量化指标 *RMSE*(均方根误差)会在本文实验部分解释。但是，随着沙漏数量呈指数增加，网络参数数量也呈指数增加，而网络性能却提升缓慢，因此，本模型采用四级沙漏网络，在保证精确度的同时，减少网络参数，提升模型性能。

因为人体分割任务和人体深度类别预测任务的本质均为像素级多分类，所以引导器网络的损失函数采用交叉熵损失(cross-entropy loss)，公式表示为：

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/12
DOI: 10.16182/j.issn1004731x.joss.19-VR0462

第 32 卷第 7 期 系统仿真学报 Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

$$loss(x,class)=-\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right)= -x[class]+\log\left(\sum_j \exp(x[j])\right) \quad (1)$$

式中：x 表示每个类别期望的分数；*class* 表示正确的类别标号。

表 1 不同任务中采用不同数量的沙漏网络单元的实验结果比较

Tab. 1 Comparison of experimental results using different numbers of hourglass network units in different tasks

沙漏数量	FSRNet[31] (PSNR/dB)	本文方法(*RMSE*/mm)
1	25.69	/
2	25.87	78.6
4	25.95	69.2
8	/	65.1

本文采用室内真实人体数据集 Human3.6M 和合成人体数据集 SURREAL 进行网络模型训练。由于数据集标注与实际任务需求有分类、数值区间以及人体画面占比等方面的差别，本文采用独特的数据预处理方式和数据增强方法。

在人体部分分割问题中，每个像素将被分类为以下 15 个类别中的一种，包括头部、身体躯干、左右大腿、左右小腿、左右上臂、左右下臂、左右手、左右脚和背景。由于 Human3.6M 数据集[23]和 SURREAL 数据集[20]以不同的标准将人体划分为 25 类，本文对其进行数据对齐，统一为上述 15 个类别。在深度类别预测问题中，需要将数据集中的绝对深度转化成相对深度。人体深度类别的划分，参照 Varol 等[20]的标准：根据正常人体所占的深度范围，以骨盆位置为中点，每隔 45 mm 作为一个深度类别区间，在骨盆前后分别取 9 个深度区间，最终将人体量化为 19 个深度区间，加上背景总共 20 类。引导器输入 RGB 图像尺寸统一规范为 256×256 像素大小。Human3.6M 数据集中人体在画面中的比重较小，难以用于训练。因此，为提高模型精准度，在训练前对数据集进行数据增强，具体方法为：使用包围盒对人体进行不同尺度的放大，使得人体在整个画面中的比重包含多个层次。为了进一步增强模型的泛化能力和抗干扰性，将训练集 RGB 图像的 3 个通道分别以不同的概率进行色调变换。

受迁移学习启发，引导器训练使用预训练和微调的方法。引导器训练分为 3 阶段：第 1 阶段用 SURREAL 数据集初始化网络的人体部分分割模块；第 2 阶段屏蔽网络前半部分，用 SURREAL 数据集初始化网络的人体深度预测部分；第 3 阶段用 SURREAL 数据集混合 Human3.6M 数据集对整个网络进行微调。网络优化方法采用随机梯度下降方法(SGD)。

2.2 人体深度修复

人体深度修复部分主要由生成器和判别器组成。由于深度类别预测结果能够精准的保留人体的局部特征，并且与真实深度存在近似线性的关系，人体深度修复部分的网络以深度类别为引导，将其作为生成器输入的一部分，从而指导网络产生更符合人体特征的结果，如图 3 所示。

在生成器中，输入包括去噪并以均值填充孔洞后的人体孔洞图和预测得到的人体深度类别特征(分辨率均为 256×256)，使用不同的卷积层将它们分别提取特征后将得到的特征串接，再将特征进行编码，得到原图 1/16 大小的特征图。然后再对该网络采用 2 层反卷积和 3 层卷积对其进行解码，从而得到与原始尺寸形同的人体深度修补图。生成网络模型仅仅降低了两次分辨率，使用大小为原始大小四分之一的卷积，其目的是为了降低最终图像的纹理模糊程度。

要使得孔洞得到平滑的修复，需要增大卷积核的感受野，使卷积核能够更好的对孔洞周围的特征进行提取。为此生成器网络模型采用扩张卷积[32]代替普通卷积，扩张率设置为 4 倍。

单使用均方差损失得到的结果会出现纹理粗糙，边缘不平滑等问题，这使得结果难以应用到进一步的研究。为此，增加判别器在辅助生成器产生

更精准的结果。判别器网络的输入为修补后的人体深度图(分辨率为 256×256)，通过卷积网络提取特征，得到 1 024 维度的向量，最终经过全连接和 Sigmoid 激活函数，将输出值控制在(0, 1)范围内，该值表示输入图像为真实图像的概率。

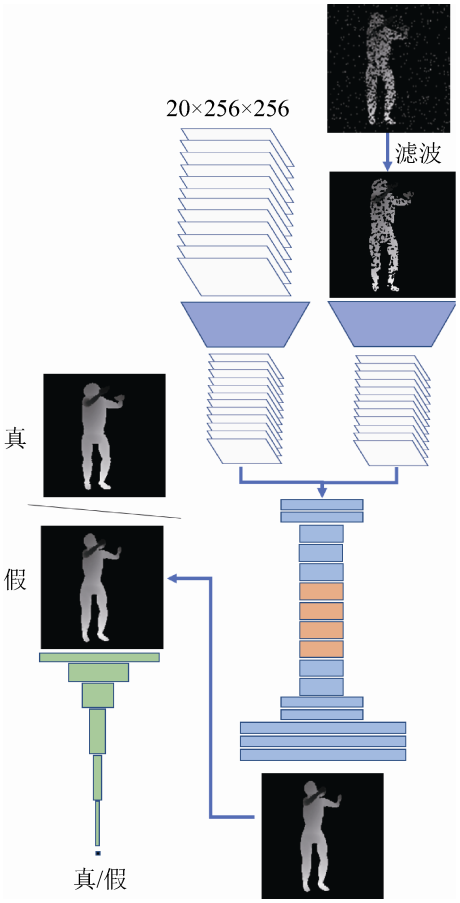

图 3 生成器与判别器网络结构图
Fig. 3 Architecture of generator and discriminator

生成器和判别器训练使用 Human3.6M 数据集。在训练之前，为数据集中的完整深度图添加随机孔洞，以模拟深度相机中的孔洞。

生成器网络使用均方差损失(MSE Loss)和对抗损失(GAN Loss)相结合的方式训练。前者有助于稳定训练，后者有助于提升结果的真实性。

均方差损失定义如下：

$$L(x,D_C)=\|M_C\odot(C(x,D_C)-x)\|^2 \tag{2}$$

式中：$\odot$ 为像素级的相乘；$\|\cdot\|$ 为欧氏范数；D_C 为人体深度类别信息；M_C 为待填补孔洞的遮罩(Mask)，M_C 由人体深度类别预测结果计算得到。

对抗损失(GAN Loss)是训练的关键部分，包括将神经网络的标准优化问题转化为最小-最大优化问题，在每次迭代中，判别器网络与生成器网络共同更新，优化方法如下：

$$\min_C \max_D \mathbb{E}[\log D(x,\mathrm{D_C})+\log(1-D(C(x,\mathrm{D}_C),\mathrm{D}_C)] \tag{3}$$

式中：D_C 为人体深度类别信息。结合 2 个损失函数，总体优化方法如下：

$$\min_C \max_D \mathbb{E}[L(x,D_C)+\alpha\log D(x,\mathrm{D_C})+\alpha\log(1-D(C(x,\mathrm{D}_C),\mathrm{D}_C)] \tag{4}$$

式中：α 为权重超参数，本模型将其设置为 0.001。本章网络采用 ADADELTA 优化算法，该算法可以自动为网络中的每个权重设置一个学习率。

在网络训练时首先使用均方差损失训练生成器，之后固定生成器参数，单独训练判别器，最后同时训练生成器和判别器。

3 实验结果与分析

3.1 人体深度类别预测

本文使用均方根误差(*RMSE*)和缩放平移不变的均方根误差(*st-RMSE*)这两个指标作为评价标准，同时也使用准确率作为参考指标。在量化时，需要将预测的深度类别结果 C_i 转换为深度值 D_i，两者之间的关系如下：

$$\hat{D}_i=t+s\cdot C_i \tag{5}$$

式中：t 为平移因子(translation)；s 为放缩因子(scale)；C 为深度函数，在此处 C 为离散值，范围是 1~20 的正整数，值越大代表对应的深度值越大。人体深度类别预测部分首先采用 *RMSE*(*RMSE* 指的是均方根误差，即测量值 $\hat{D}$ 和真实值 D 之间的误差)来作为度量标准。由于在标记训练数据时，本章采用 45 mm 作为分类标准，即每 45 mm 作为一个深度类别，因此在 *RMSE* 计算中，使得 s=45，如公式(6)所示。

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/12
DOI: 10.16182/j.issn1004731x.joss.19-VR0462

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(t+45\cdot C_i-D_i)^2}{n}} \tag{6}$$

在 *RMSE* 的计算中，设置 s 为 45 mm。为了能够在定量分析时体现当前结果的放缩不变性和平移不变性，实验引入了 *st-RMSE* 指标(*st-RMSE* 为：在放缩因子和平移因子都是最优的情况下，测量值 $\hat{D}$ 和真实值之间的误差)。在 *st-RMSE* (scale translation invariant *RMSE*)计算中，s 为最优放缩(scale)因子，通过线性拟合求得，如公式(7)所示。

$$st\text{-}RMSE=\sqrt{\frac{\sum_{t=1}^{n}(t+s\cdot C_i-D_i)^2}{n}} \tag{7}$$

准确率定义为分类正确的像素与总体像素之比。

在 Human3.6M 数据集上使用 *RMSE*、和准确率对本文方法进行量化，结果如表 2 所示。

表 2 不同方法在 Human3.6M 数据集上的人体深度预测结果对比

Tab. 2 Comparison of human depth prediction results on Human3.6M dataset by different methods

方法	*RMSE*/mm	*st-RMSE*/mm	准确率%
Varol et al.	90.0	67.1	87.63
本文方法	69.2	51.7	92.89

由于结合了人体部分解析的特征，模型能够更好的理解人体，相比 Varol 等的方法，本文提出的方法能够保持人体的完整性，结果也更符合人体的结构和特性；沙漏网络的改进也使得结果从细节上更加精确。结果展示如图 4 所示，从左到右每列分别为：RGB，标注的相对人体深度类别(Ground Truth)，Varol 等的方法预测的相对人体深度，本文方法预测的相对人体深度。

3.2 人体深度修复

人体深度修复的结果在 Human3.6M 数据集上使用 *RMSE* 进行量化分析，公式为：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(I_i-D_i)^2}{n}} \tag{8}$$

式中：I 为修复后的人体深度图。

分析结果如表 3 所示。泊松融合[33]为图像领域经典的方法，本文尝试将泊松融合加入网络的后续处理。结果显示，泊松融合在修复 Human3.6M 数据集人造孔洞时表现出色。

图 4 人体深度预测阶段结果

Fig. 4 Results of human depth prediction

表 3 不同方法在 Human3.6M 数据集上的人体深度修复结果对比

Tab. 3 Comparison of human depth restoration results on Human3.6M dataset by different methods

方法	峰值信噪比 (PSNR/dB)	均方根误差 (*RMSE*/mm)
Satoshi 等[30]	19.91	130.71
均方差+泊松融合	29.75	37.91
引导对抗网络	29.03	38.80
引导对抗网络+泊松融合	30.31	36.66

实验中，将 Satoshi 等的方法、“均方差+泊松融合”、“引导对抗网络”和“引导对抗网络+泊松融合”进行结果对比，结果如图 5 所示，从上到下每行依次为 RGB 图像、标注的绝对人体深度(Ground Truth)、人工添加孔洞和噪声后的人体深度图、Satoshi 等方法的结果、“均方差+泊松融

合”的结果、引导对抗网络的结果、“引导对抗网络+泊松融合”的结果。其中，Satoshi 等的方法可以代表彩色图像修复领域最高水平，但是其方法需要预知孔洞的位置才可以对其进行修复。在使用 Satoshi 等的方法时，孔洞位置由本文方法人体深度类别预测结果计算得到。

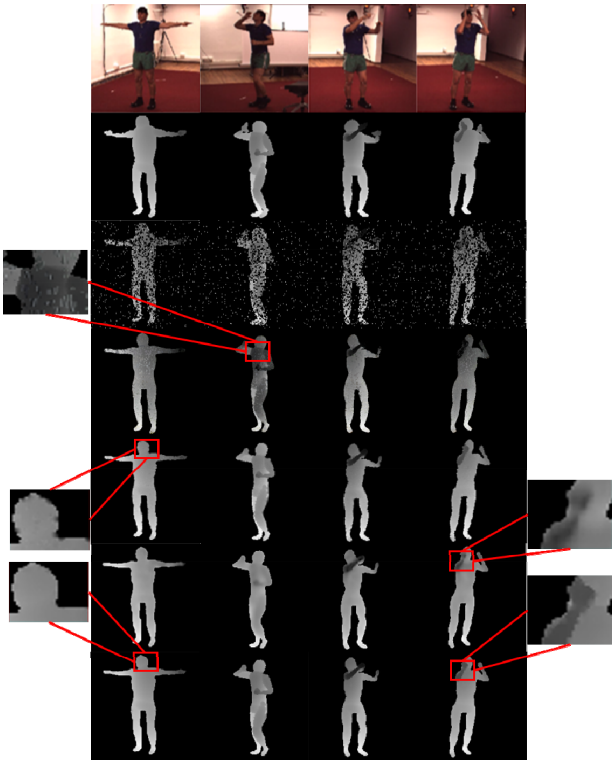

图 5　在 Human3.6M 数据集上的人体深度修复阶段结果
Fig. 5　Result of human depth restoration on Human3.6M

由图 5 可见，相比于 Satoshi 的方法，“引导对抗网络+泊松融合”方法能够更加真实的还原深度数据；相比“均方差+泊松融合”方法，“引导对抗网络+泊松融合”方法能够解决其边缘锯齿的问题，生成更精确的结果。在人体出现肢体交叠时，泊松融合能够使得修复结果变得更加清晰；在人体肢体舒展时，泊松融合没有出现明显效果。

用小型深度相机 RealSense 进行实验，结果如图 6 所示，前 2 行为 RealSense 采集到的 RGB 图像和深度图(已对齐)，第 3 行为人体部分分割结果，第 4 行第 5 行分别为 Varol 等的方法和本文方法的人体深度类别预测结果，第 6 行为引导对抗网络的人体深度修复结果，第 7 行为“引导对抗网络+泊松融合”的人体深度修复结果。

本文方法在实验过程中使用的 GPU 为 NVIDIA 1080TI，实验环境为 cuda 9.1 和 cudnn 7，对于每次前向运算(forward)需要时间大约 60 ms。

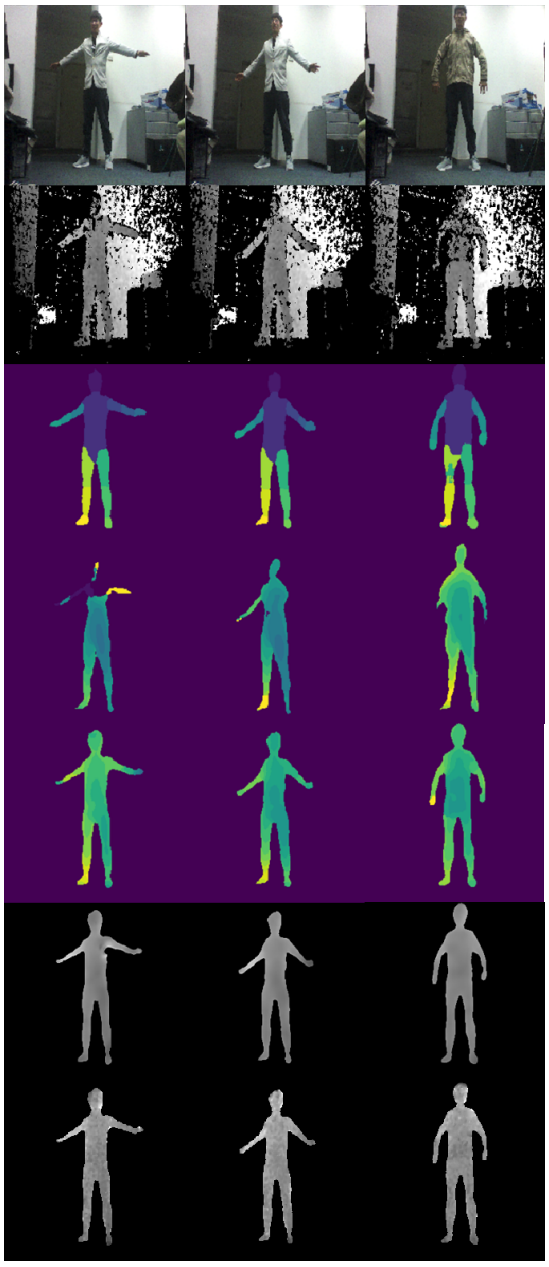

图 6　在 RealSense 采集到的数据集上的人体深度修复结果
Fig. 6　Results of human depth restoration on dataset captured by RealSense

由图 6 可见，在实际应用中，引导对抗网络在人体深度预测上优于 Varol 等的方法，也出色地

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/12
DOI: 10.16182/j.issn1004731x.joss.19-VR0462

第 32 卷第 7 期 系统仿真学报 Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

解决了人体深度修复问题。但是，令人意想不到的是，“引导对抗网络+泊松融合”的结果粗糙，泊松融合的作用适得其反。本文就这一结果进行如下分析。

在实际的应用中，深度相机拍摄的人体往往有复杂着装，所以实际拍摄到的深度图会保留服装材质、形状、褶皱的深度信息，造成深度图的表面出现粗糙的斑点；而 Human3.6M 数据集中的深度图是通过复杂的传感器设备获取，并经过作者的特殊处理，其表面光滑无服装特征。由于泊松融合[34]能够保留原始深度图中的有效信息，所以在使用实际深度相机实验时，原始深度图中粗糙的服饰特征也会被保留；而用人体数据集则不存在服装的干扰。因此，在人体数据集和深度相机采集到的数据集中，泊松融合才会出现相反的表现。

尽管在人体数据集的实验中，泊松融合[34]可以很好地处理人体肢体交叠情形的修复精准度，但在实际的人体测量和人体建模应用中，为了提高测量精准度，会要求人体摆出正规的姿势，不会允许出现肢体交叠情形。

实验结果以及上述分析可见，在实际应用中，引导对抗网络能够在保留人体特性的同时去除服装的干扰，有效修复人体深度图中的孔洞。

4 结论

本文提出了引导对抗网络，有效解决了移动端小型深度相机的孔洞问题。实验证明，本论文提出的方法，在现有的包含深度信息的人体数据集上有比较出色的表现，无论是从视觉上还是实际数据质量上，均取得了很好的效果。实际应用时，能够在保留人体特性的同时去除服装的干扰，有效修复人体深度图中的孔洞。

在未来的工作中，我们会进一步将引导对抗网络到普通移动设备上的人体建模和非接触式人体测量，以优质可靠的人体深度数据，推进虚拟试衣、人机交互、游戏、影视等领域的发展。

参考文献：

[1] Tukey J W. Exploratory Data Analysis (preliminary ed.)[M]. Reading MA: Addison-Wesley, 1971.

[2] Bertalmio M, Sapiro G, Caselles V, et al. Image Inpainting[C]. In ACM Transactions on Graphics. USA: ACM Press/Addison-Wesley Publishing Co., 2000, 417-424.

[3] Barnes C, Shechtman E, Finkelstein A, et al. Patch-Match: A Randomized Correspondence Algorithm for Structural Image Editing[J]. ACM Transactions on Graphics (S0730-0301), 2009, 28(3): 24-35.

[4] Darabi S, Shechtman E, Barnes C, et al. Image Melding: Combining Inconsistent Images using Patch-based Synthesis[J]. ACM Transactions on Graphics (S0730-0301), 2012, 31(4): 82-92.

[5] Köhler R, Schuler C, Schölkopf B, et al. Mask-specific inpainting with deep neural networks[C]. In German Conference on Pattern Recognition. German: Springer, Cham, 2014: 523-534.

[6] Ren J S J, Xu L, Yan Q, et al. Shepard Convolutional Neural Networks[C]. In Conference on Neural Information Processing Systems. New York: Curran Associates, Inc., 2015: 901-909.

[7] Xie J Y, Xu L L, Chen E H. Image Denoising and Inpainting with Deep Neural Networks[C]. In Conference on Neural Information Processing Systems. New York,NY: Curran Associates, Inc., 2012: 341-349.

[8] Pathak D, Krähenbühl P, Donahue J, et al. Context Encoders: Feature Learning by Inpainting[C]. In IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2536-2544.

[9] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets[C]. In Conference on Neural Information Processing Systems. New York: Curran Associates, Inc., 2014: 2672-2680.

[10] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]. European Conference on Computer Vision. Cham: Springer, 2016: 483-499.

[11] Dollár P, Wojek C, Schiele B, et al. Pedestrian Detection: A Benchmark[C]. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 304-311.

[12] Pishchulin L, Andriluka M, Gehler P, et al. Poselet conditioned pictorial structures[C]. In CVPR. Portland, Oregon: IEEE, 2013: 588-595.

[13] Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts[C]. In CVPR. Providence, RI: IEEE, 2011. 1385-1392.

[14] Chen X, Yuille A L. Articulated pose estimation by a graphical model with image dependent pairwise relations[C]. In NIPS. New York: Curran Associates, Inc., 2014. 1736-1744.

[15] Chu X, Ouyang W, Li H, et al. Structured feature learning for pose estimation[C]. In CVPR. Las Vegas, NV: IEEE, 2016. 4715-4723.

[16] Zhou X, Huang Q, Sun X, et al. Towards 3d human pose estimation in the wild: a weakly-supervised approach[C]. In IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 398-407.

[17] Yang W, Ouyang W, Wang X, et al. 3D Human Pose Estimation in the Wild by Adversarial Learning[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018, 5255-5264.

[18] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA: IEEE, 2015: 3431-3440.

[19] Zhao J, Li J S, Cheng Y, et al. Understanding Humans in Crowded Scenes: Deep Nested Adversarial Learning and A New Benchmark for Multi-Human Parsing[C]. In MM '18 Proceedings of the 26th ACM international conference on Multimedia. New York: Association for Computing Machinery, 2018: 792-800.

[20] Varol G, Romero J, Martin X, et al. Learning from synthetic humans[C]. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Venice: IEEE, 2017: 109-117.

[21] Jacques J J C S, Jung C R, Musse S R. Skeleton-based human segmentation in still images[C]. 19th IEEE International Conference on Image Processing, Orlando, FL: IEEE, 2012: 141-144.

[22] Rother C, Kolmogorov V, Blake A. "GrabCut": interactive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics (S0730-0301), 2004, 23(3): 309-314.

[23] Ionescu C, Papava D, Olaru V, et al. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments[J]. IEEE transactions on pattern analysis and machine intelligence (S0162-8828). 2014, 36(7): 1325-1339.

[24] Loper M, Mahmood N, Romero J, et al. SMPL: A Skinned Multi-Person Linear Model[J]. ACM Transactions on Graphics (S0730-0301), 2015, 34(6): 248.

[25] Ma X, Geng Z, Bie Z. Depth estimation from single image using cnn-residual network[Z].

[26] Chen W, Fu Z, Yang D, et al. Single-image depth perception in the wild[C]. Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2016: 730-738.

[27] Liu B, Gould S, Koller D. Single image depth estimation from predicted semantic labels [C]. Computer Vision and Pattern Recognition(CVPR), San Francisco : IEEE, 2010: 1253-1260.

[28] Liu W, Chen X, Yang J, et al. Robust Color Guided Depth Map Restoration[J]. IEEE Transactions on Image Processing (S1057-7149), 2017, 26(1): 315-327.

[29] Hays J, Efros A A. Scene Completion Using Millions of Photographs[J]. ACM Transactions on Graphics (S0730-0301), 2007, 26(3): 4-12.

[30] Satoshi Iizuka, EDGAR SIMO-SERRA, HIROSHI ISHIKAWA. Globally and Locally Consistent Image Completion[J]. In ACM Transaction on Graphics (S0730-0301), 2017, 36(4): 1-14.

[31] Chen Y, Tai Y, Liu X M, et al. FSRNet: End-to-End Learning Face Super-Resolution With Facial Priors[C]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT: IEEE, 2018: 2492-2501.

[32] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[C]. In International Conference on Learning Representations. San Juan, Puerto Rico: Computational and Biological Learning Society, 2016.

[33] Pérez P, Gangnet M, Blake A. Poisson Image Editing[J]. ACM Transactions on Graphics (S0730-0301), 2003, 22(3): 313-318.

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/12
DOI: 10.16182/j.issn1004731x.joss.19-VR0462