DOI: 10.3785/j.issn.1008-973X.2019.01.013

基于生成对抗网络的图像恢复与 SLAM 容错研究

王凯，岳泊暄，傅骏伟，梁军
(浙江大学 控制科学与工程学院，浙江 杭州 310058)

摘 要：为了提高即时定位与地图构建(SLAM)系统的容错能力，在经典图像生成网络 Pix2Pix 的基础上，逐步添加深度估计网络和深度信息的输入、基于 STN 网络的图像重建损失以及基于图像修复网络的图像补全损失 3 个方面的改进. 结合双目图像的耦合关系，通过挖掘和融合多种信息，增大了信息的利用率，提高了模型的图像生成效果. 提出将生成对抗网络(GAN)技术与 SLAM 容错场景相结合，直接实现了感知端的容错. 在 KITTI 和 Cityscapes 数据集上进行实验，验证了改进模型的有效性. 将模型生成的图像用于双目视觉系统的重建，验证了容错思想的可行性.

关键词：图像生成网络；容错；即时定位与地图构建(SLAM)；图像恢复；Pix2Pix
中图分类号：TP 183　　**文献标志码：**A　　**文章编号：**1008−973X(2019)01−0115−11

Image restoration and fault tolerance of stereo SLAM based on generative adversarial net

WANG Kai, YUE Bo-xuan, FU Jun-wei, LIANG Jun
(*College of Control Science and Engineering*, *Zhejiang University*, *Hangzhou* 310058, *China*)

Abstract: The classical Pix2Pix network was modified in order to promote the capacity of fault tolerance of simultaneous localization and mapping (SLAM) system. The network was gradually added to depth estimation network and its depth information, image reconstruction loss based on STN network and image inpainting loss based on image inpainting network. Information was mined based on the coupling of stereo images and merged to utilize information usage and promote model performance. Then generative adversarial net (GAN) and SLAM were combined, and the fault tolerance in the sensing level was directly realized. Experiments were performed on KITTI and Cityscapes dataset in order to prove the effectiveness of the improvement. The generated images and original images were both fed as inputs of stereo SLAM system. Results showed that the fault tolerance idea was approachable.
Key words: image generation network; fault tolerance; simultaneous localization and mapping (SLAM); image restoration; Pix2Pix

自动驾驶场景中当某一目视角图像出现问题时如何根据某一目图像进行目标视角图像的恢复是一个新的研究方向. 该方向在位姿重建[1]及深度场景理解等方面有很多应用[2]. 基于计算机视觉的 3D 环境感知主要有以下 2 种方式：基于单目的即时定位与地图构建(monocular SLAM)[3-4]和基于双目的即时定位与地图构建(stereo SLAM)[5-6]. 大部分这类方法假设多个视角的观察数据能够获取，比如多目的视觉图像信息或者是不同时刻不同位置对于某一物体的视觉图像. 该类方法的缺点是如果某一视角图像出现问题(遮挡、强光或者设备问题没采集到信号)，则不能很好地重建

收稿日期：2018−01−10.　　**网址：**www.zjujournals.com/eng/fileup/HTML/201901013.htm
基金项目：国家自然科学基金资助项目(U1664264, U1509203).
作者简介：王凯(1993—)，男，硕士生，从事计算机视觉的研究. orcid.org/0000-0002-4349-6486. E-mail: kaiwang1@zju.edu.cn
通信联系人：梁军，男，教授. orcid.org/0000-0003-1115-0824. E-mail: jliang@zju.edu.cn

场景信息. 为了克服基于多个视角图像合成场景的限制, 最近有大量的工作结合机器学习和深度学习, 研究基于单视角的新视角图像生成以及深度估计问题[7-11]. 该类方法通过有监督学习的方式, 针对大量有标签的数据集进行训练, 得到新视角生成模型并生成可观的新视角图像, 从而获取深度场景信息. 根据训练数据集的不同, 该类有监督学习的方法可以分为以下 2 类: 1)以单视角图像信息作为输入, 深度点云信息作为输出, 通过多层网络进行映射学习, 将学习到的深度信息用于多视角图像的生成[12-14]; 2)以源视角图像信息作为输入, 目标视角的图像作为输出, 通过多层网络进行映射学习, 将学习到的图像用于场景图像及深度图像的重建[15-19].

本文结合 Godard 等[20]的工作, 提出新的图像生成网络. 首先利用 Godard 提出的深度估计网络生成目标视角的深度图, 同时重建目标视角图像, 得到一个中间结果; 然后利用图像补全网络完成之前中间结果的修正.在生成重建图像的过程中, 利用深度图像的梯度掩码图. 在得到良好的目标视角图像后, 将 2 个视角的图像作为 ORB_SLAM 的输入, 生成较完备的场景信息及深度信息. 本文在经典的 Pix2Pix 网络[21]的基础之上, 提出 3 点改进: 1)添加深度先验信息; 2)添加目标视角图像重建损失; 3)添加深度梯度掩码损失. 本文通过在 KITTI 和 Cityscapes 数据集上的实验, 证实了改进的有效性.

1 相关工作

最近有很多研究人员开展了基于单目的图像重建以及深度估计的相关工作. Jaderberg 等[7]提出空间转换网络(spatial transform network, STN), 网络在使用时可以显式地指定图像的转换参数, 能够插入到常用的卷积神经网络架构中, 使得神经网络能够基于特征层的数据, 进行图像转换. 该网络具有尺度、旋转、扭曲等不变性, 是深度学习用于图像变换的早期工作, 为后面的部分工作提供了工作基础. Tatarchenko 等[8]提出能够从单一视角图像生成任意视角图像的网络. 网络输入是单一视角图像及欲观察角度, 输出是指定角度的生成图像. 该网络学习到某一物体类的隐式 3D 表达, 能够将该类知识迁移到新的实例中去. 该网络在生成渲染图像的同时, 可以生成与之对应的深度信息, 即 3D 场景信息. 网络只能针对特定物体进行训练和预测. Zhao 等[9]解决了从单一视角图像生成多视角图像的问题. 本文提出新的生成模型 VariGAN, 该模型结合了变分推断及生成式网络的优势. 网络首先通过变分推断对物体整体的外貌进行建模, 包括形状和颜色; 然后通过生成对抗网络(GAN)产生纹理更加细致的图像, 整个过程以一种递进式的方法, 生成从粗糙逐渐精细的图像. Zhou 等[10]提出 appearance flow 的概念, 认为视角与视角之间的变换可以表达为 appearance flow, 即像素的转移方向和大小. 通过训练 CNN 网络能够更好地预测像素的转移流场, 基于该流场对图像进行相应的变换, 得到新的视角图像. 实验证明, 利用该方法得到的新的生成图像, 在图像感知评估指标上高于以前的方法. Park 等[11]在 Zhou 等的基础上提出若干点改进. 该网络首先将输入图像上可以看到的图像通过空间转换网络变换到输出图像上, 剩下源图像不存在的图像部分通过图像补全来完成. 图像补全采用 GAN 网络完成, 损失函数采用对抗损失和特征感知损失. 该网络在图像的生成质量上取得了很好的效果, 缺点是图像的转换角度是显式指定的, 这在自动驾驶场景中不适用. Godard 等[20]完成了利用单一视角图像生成深度图像的工作; 主要提出新的损失函数, 即对极几何损失及图像重建损失. 对极几何损失主要用于约束左、右图像的一致性, 图像重建损失用于约束生成图像与真实图像的一致性. 该网络在深度图像的预测上取得了很好的效果, 实现了多视角图像的重建, 但是图像容易发生个别物体位置不精确的问题.

2 针对 Pix2Pix 与深度估计网络的损失函数改进

2.1 深度估计网络

深度估计问题定义如下: 给定源视角图像 I, 为了方便, 输入采用双目中的左视角图像 I^{l}, 网络的目标是学习到一个映射 f, 预测像素级别的深度场景信息 $d=f\left(I^{\mathrm{l}}\right)$. 已存的部分学习算法将该问题看成有监督学习问题: 输入是单一视角的图像, 输出是目标深度图信息. 在实际场景中, 获得如此大量的深度图信息比较困难, 信息也不准确(即使昂贵的激光扫描设备, 也会因为运动和反射导致测量不精确). 深度估计网络将深度估计

问题看成图像重建问题. 如果能够学到一个函数将一个视角的图像映射成另一个视角的图像, 那么可以认为该函数学习到了场景中的某种 3D 信息. 基于有监督学习的图像重建问题定义如下：在训练阶段, 给定输入 I^{l} 和 I^{r}, 分别对应于双目中同一时刻经过矫正的左视角彩色图像和右视角彩色图像；函数 g 根据学习到的像素流场 d^{r}, 对源视角图 I^{l} 进行像素级转换, 得到右视角的重建图像 $\bar{I}^{\mathrm{r}}$, 即 $g\left(I^{\mathrm{l}},d^{\mathrm{r}}\right)=\bar{I}^{\mathrm{r}}$. 根据右视角图, 可以得到左视角图的重建, 即 $g(I^{\mathrm{r}},d^{\mathrm{r}})=\bar{I}^{\mathrm{l}}$. 其中像素流场 d 是要学习到的左、右视角图像的视差图. 根据对极几何可知, 视差图和深度图有如下关系：

$$Z=fb/d. \tag{1}$$

式中：b 为双目相机的基线, Z 为物体的深度, f 为相机的焦距. 根据式(1), 通过获得 d, 可以获得相应的相对深度信息.

深度估计网络通过生成视差图, 然后利用视差图将源视角图像变换到目标视角图像, 构建与真实目标视角图的重建误差. 图像的生成函数 g 采用 STN 网络[7]. 该网络将预计生成的像素点通过视差图 d 反向映射到源视角图中的某坐标, 取该坐标的像素作为目标点值. 对于没有刚好映射到整数坐标的像素, 采用双线性插值采样. 双线性插值公式可导, 能够在神经网络的学习中传递误差信息.

深度估计网络模型框架采用经典的编码器-解码器结构, 具体如图 1 所示.

模型在训练阶段输入左视角图像和右视角图像, 测试阶段只输入左视角图像. 模型在构建重建误差的同时, 将视差图训练出来. 视差图生成网络包含短路连接[22], 增加了编码器部分和解码器部分的信息共享.

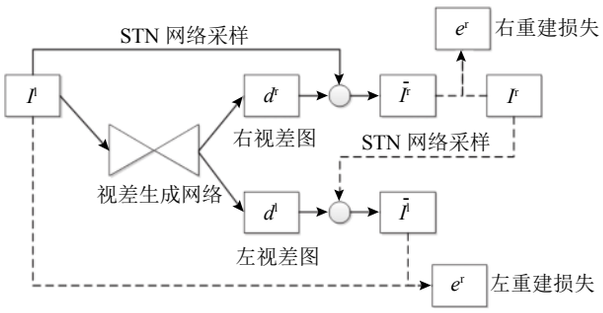


图 1　带有图像重建损失的深度估计网络结构图

Fig.1　Architecture of depth estimation network with image reconstruction error

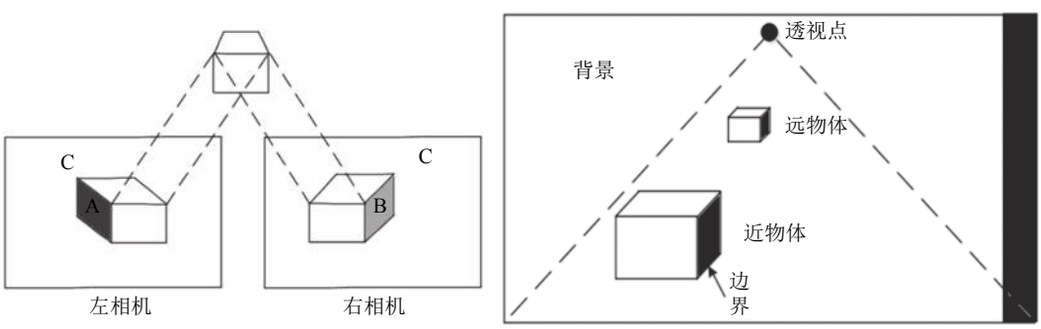
网络的损失函数采用多尺度损失. 将每层输出尺度 s 的损失求和, 得到总的损失 $C=\sum_{s=1}^{4}C_s$, 每个尺度的损失由以下几个损失构成：

$$\begin{aligned}C_{\mathrm{s}}=&\alpha_{\mathrm{recon}}\left(C_{\mathrm{recon}}^{\mathrm{l}}+C_{\mathrm{recon}}^{\mathrm{r}}\right)+\\&\alpha_{\mathrm{lr}}\left(C_{\mathrm{lr}}^{\mathrm{l}}+C_{\mathrm{lr}}^{\mathrm{r}}\right)+\alpha_{\mathrm{ds}}\left(C_{\mathrm{ds}}^{\mathrm{l}}+C_{\mathrm{ds}}^{\mathrm{r}}\right).\end{aligned} \tag{2}$$

式中：α_{recon}、α_{lr} 和 α_{ds} 为需要调节的参数；C_{recon} 为重建损失, 约束生成图像和真实目标图像的一致性；C_{ds} 为视差图的平滑约束损失；C_{lr} 约束左、右视差图的一致性. 损失的具体计算方法可以参见文献[15].

2.2　改进的图像补全网络

2.1 节的深度估计网络是新视角重建网络的第 1 个子网络. 该网络用于获取图像的视差信息, 根据源视角图像双线性采样, 得到目标视角图像. 经过 STN 视角转换得到的网络存在不够全面的问题. 新视角图像中的像素元素可以看成以下几种类型的集合：1)源视角图里的像素在目标视角图中保持可见, 仅仅是位置发生了相应的变换；2)源视角图中的像素在目标视角图中消失, 变得不可见；3)源视角图中不可见的像素在目标视角图中出现, 3 种类型如图 2(a)所示.



(a) STN 网络的中间结果　　(b) 距离远近对物体成像影响示意图

图 2　成像与物体距离关系的示意图

Fig.2　Relation between formation of images and distance

图 2（a）中左、右框为左、右两个相机所拍摄到的 2 个视角图，图像内白色部分为类型 1）的像素块，深灰色为类型 2）中的像素块，淡灰色部分为类型 3）中的像素块. 经过分析可知，在成像的过程中，物体离相机的距离越远，形成的不同视角图之间的偏移越小；距离越近，不同视差造成的偏移越大，因此源视角生成的目标视角图像中会有更大的空白需要补全和填充. 另外，A、B 区域通常发生在图像中物体的边界部分. 因为物体和背景的分割线通常是物体的边界，不同物体离相机焦点的距离通常不同，于是当图像发生视角转换时，不同距离会带来不同的视差变化，在边界处发生边界的堆挤或者割裂. 这种堆挤和割裂分别形成了类型 2）和 3）中的像素. 距离越近，这种效应越明显，生成另一视角形成的空隙越大. 这种效果如图 2（b）所示.

根据上述讨论，改进图像补全网络模型，添加了若干改进并形成了多种模型，不同模型存在部分的变化.

2.2.1 *深度估计网络与 Pix2Pix 结合* 该模型结合深度估计网络与 Pix2Pix 网络[21]，将原本新视角图像生成的 3 个子任务分成以下 2 步：1）利用深度估计网络完成 2 个视角中共有像素的复制和转移以及源视角图像中部分像素的去除，得到目标视角待补全的中间结果；2）利用 Pix2Pix 的图像生成能力，完成源视角中不存在部分像素区域的补全. 网络将图像生成任务与像素转移任务解耦，同时提高每个子任务模型的性能. 网络之间只共享待补全的视角图，训练时并非整体训练，而是分步训练. 参数更新规则如下：

$$\left.\begin{aligned} W_{\text{mono}} &= W_{\text{mono}} - \rho_{\text{mono}} \sum_{i=1}^{M} \frac{\partial L_{\text{adv}}^{\text{mono}}\left(X^i, Y^i\right)}{\partial W_{\text{mono}}}, \quad \text{step} \leqslant h; \\ W_{\text{D}} &= W_{\text{D}} - \rho_{\text{D}} \sum_{i=1}^{M} \frac{\partial L_{\text{adv}}^{\text{D}}\left(X^i, Y^i\right)}{\partial W_{\text{D}}}; \quad \text{step} > h; \\ W_{\text{G}} &= W_{\text{G}} - \rho_{\text{G}} \sum_{i=1}^{M} \frac{\partial L_{\text{adv}}^{\text{G}}\left(X^i, Y^i\right)}{\partial W_{\text{G}}}. \quad \text{step} > h. \end{aligned}\right\} \tag{3}$$

式中：下标 mono、D 与 G 分别表示视差生成网络、Pix2Pix 中的判别网络以及生成网络；ρ 为学习率；M 为随机批次的大小；W 为网络权重与偏置参数的集合；h 为深度估计网络的训练阈值，大于该阈值后，深度估计网络的参数不再更新，只更新 Pix2Pix 网络的参数. 参数的更新规则采用随机梯度下降的方式. 具体的网络结构如图 3 所示. 图中，虚线部分表示网络学习的第 1 阶段，即深度估计网络的参数更新；实线部分表示训练的第 2 阶段，是 Pix2Pix 网络的参数更新.

2.2.2 *带有图像一致性损失的 Pix2Pix* 深度估计网络不仅重构了目标视角的图像，而且重建了源视角图像，因此一次深度估计网络的测试推理可以产生 2 个待补全的视角图像. 后续的图像补全网络在训练时并非仅能够对目标视角图像进行重建，也可以对源视角图像进行重建. 于是，图像补全网络的训练数据量扩大了一倍. 图像补全网络的参数在训练这 2 类图像时共享.

图像在训练时补全网络，针对源视角图像和目标视角图像 2 个图像进行重建. 受到深度估计网络视差图一致性的启发，图像补全网络生成的 2 个视角图像应保持一致性的约束. 该网络的图像一致性损失公式表示如下：

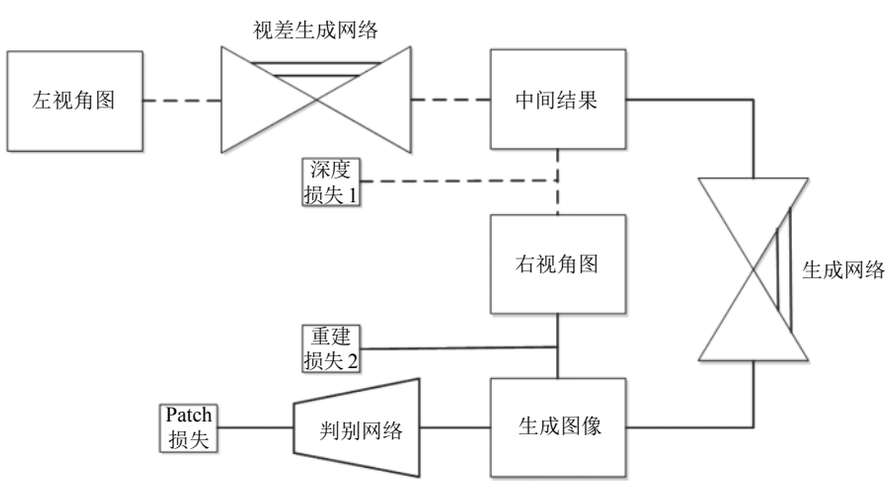


图 3 深度估计网络与 Pix2Pix 网络结合

Fig.3 Combination of depth estimation network and Pix2Pix network

$$L_{\mathrm{lr-image}} = \|c(I_{\mathrm{r}}) - T(c(I_{\mathrm{l}}), d_{\mathrm{r}})\|_2. \tag{4}$$

式中：c 表示图像补全网络，I_{r} 和 I_{l} 为待补全的左、右视角图像，T 为根据源视角及视差图生成目标视角的采样操作. 衡量图像之间的相似性可以采用特征层的损失、SSIM 等[23]衡量指标. 为了方便，在实现时采用最简单的二范数，衡量重建的目标视角图与真实的目标视角图之间的相似性. 新的训练方式以及损失函数如图 4 所示.

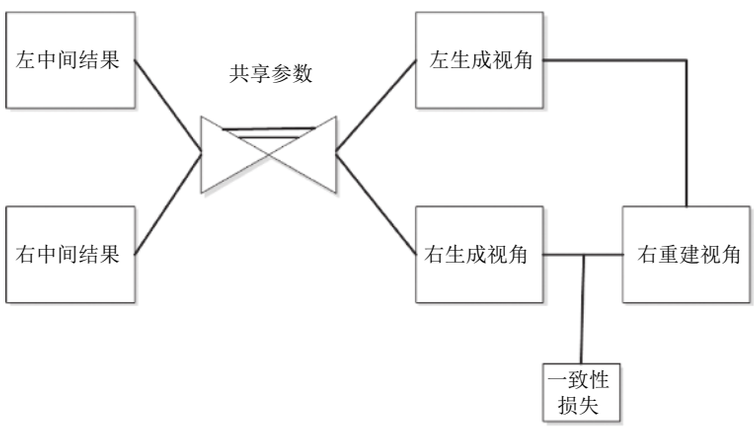


图 4　带有一致性损失的训练框架

Fig.4　Training procedure of network with coherence loss

2.2.3　带有掩码和深度特征的图像补全网络　根据 2.1 节的讨论可知，视差图像代表了图像中像素的深度信息，若视差图中的一块像素集合属于某一个物体，则该像素集合的数值差别不大；若一块像素集合和另一块像素集合的数值差别大，则这两个像素集合很可能代表不同的物体. 视差图像包含丰富的物体分割信息. 这样的物体分割信息属于图像的语义信息，若添加这样的语义信息作为图像补全网络的先验信息，则对于图像的生成会有很大的帮助. 根据上述视角缺失块的讨论可知，离相机的远近决定缺失像素块的大小，并且缺失像素的区域通常出现在物体与背景的分割线处.可以将该补全网络看成是图像修复问题[24]，掩码是通过深度图像作梯度运算获得的边缘掩码信息. 基于以上分析，可以对图像补全网络进行下述 2 点改进. 1)将图像的深度语义信息添加到输入图像的通道层；2)通过深度信息提取边缘掩码，修改图像重建损失，具体修改如图 5 所示.

通过拉普拉斯算子的卷积操作，得到边缘信息响应图. 图 5 中，边缘信息的响应变大，非边缘区域的信息被抑制. 将得到的边缘响应图通过阈值限制，获得掩码信息. 掩码求取公式如下：

$$I_{\mathrm{M}} = \left[1 - \mathrm{Int}\left(I_{\mathrm{Lap}} \geqslant T_{\mathrm{m}}\right)\right] F_{\mathrm{r}}. \tag{5}$$

式中：I_{Lap} 为边缘信息响应图；T_{m} 为筛选阈值响应的参数，通过实验确定；Int 函数将逻辑变量转换成整数型变量，即 0、1 变量；F_{r} 为和图像大小相同的掩码矩阵，因为从视角的变换，会在生成的图像边缘处形成缺失区域（见图 2(b)右边的黑色区域），该掩码将边缘缺失区域过滤掉. 式(8)

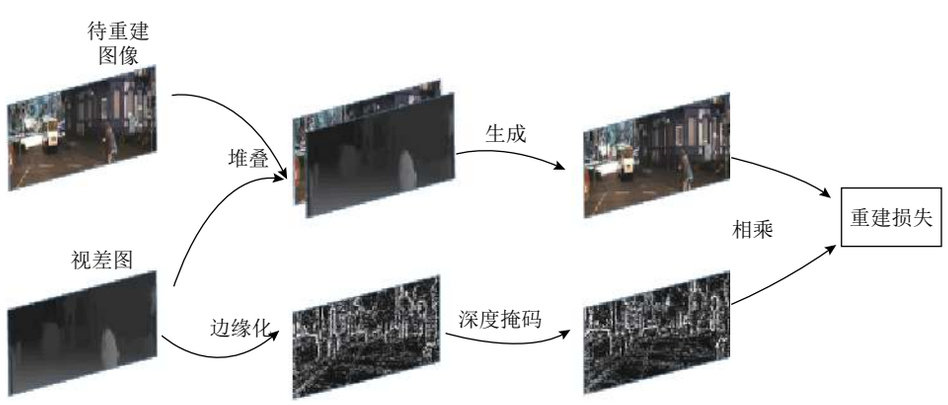


图 5　添加深度信息与梯度掩码示意图

Fig.5　Modifications with depth information and gradient mask

可以通过边缘信息响应图，得到统一的掩码矩阵. 掩码矩阵表示源图像中存在的像素区域为 1，需要根据周围的像素信息推理的空白区域为 0. 将该矩阵和图像补全网络的生成结果取掩码，得到新的图像重建损失，公式如下：

$$L_{\mathrm{rec}}(x)=\left\|I_{\mathrm{M}}\odot\left[x-G((1-I_{\mathrm{M}})\odot x)\right]\right\|_2. \tag{6}$$

3 实验分析

KITTI 数据集的训练样本为 29 000 个，测试样本 12 223 个. Cityscapes 数据集训练样本有 22 973 个，测试样本有 1 525 个. 2 个数据的参数设置一致. 4 个模型的 L_1 重建损失系数为 0.000 25，GAN 的损失系数为 1.0，L_1 损失包括重叠区域损失和重建区域的损失，损失的系数分别为 5.0 和 1.0. 模型生成器的学习率是 0.000 2，前 3 个模型判别器的学习率是 0.000 27，第 4 个模型判别器的学习率是 0.000 3. 带有梯度掩码信息的模型，阈值设置为 0.001.

3.1 KITTI 数据集实验

4 个模型的实验结果如图 6 所示. 从 KITTI 实验结果可以看出，Pix2Pix 的基准模型生成的质量最差，模型 2（mono+Pix2Pix）网络的生成质量相对于 Pix2Pix 基准模型有很大提升. 比如在第 1 个测试样例中（从左往右），从汽车左边的强光区域的生成状态来看，Pix2Pix 将接近矩形的强光区生成为椭圆形的区域. 后续的改进都极大地改进了该问题. 在第 3 个样例中，图像的右边区域实际上是茂密的绿色树叶，在 Pix2Pix 的生成图形中出现了淡绿色的模糊块，其他模型生成较合理. 说明本文的模型相对于经典的图像生成模型 Pix2Pix，有了很大的改进.

在上述的生成图像中，后续改进的图像质量相比 Pix2Pix 模型有了很大的提高，这通过直接观察生成图像的很多细节点可以看出. 改进模型之间的对比和提高很难直接观察出来，因此计算图像之间的 L_1 误差图，并转换为灰度图，实验结果如图 7 所示.

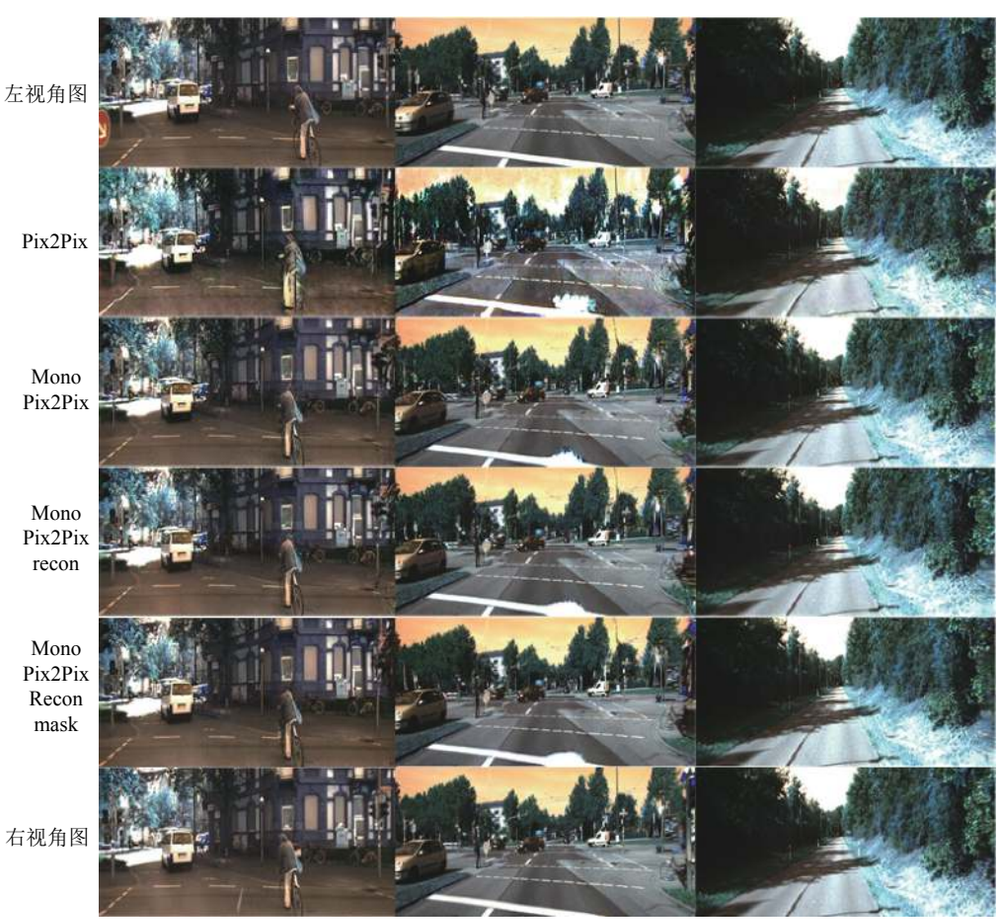


图 6 KITTI 数据集下不同模型实验结果对比图

Fig.6 Experiment results of different models on KITTI dataset

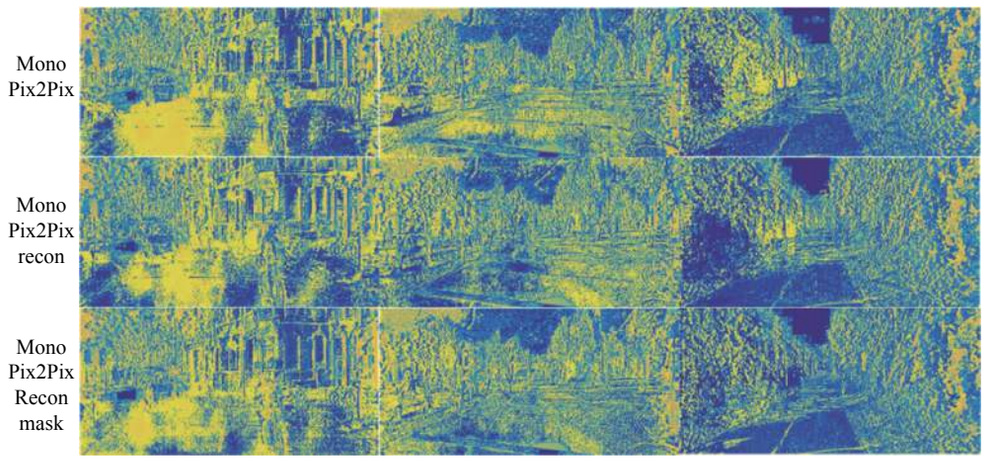


图 7　多种改进模型在 KITTI 数据集上的实验对比

Fig.7　Contrasts of experiment results between modified models on KITTI dataset

通过实验结果可以看出, 模型 3(mono+Pix2Pix+recon)较模型 2(mono+Pix2Pix)存在更大的黑色区块, 尤其是样例 1 和样例 3 可以看出, 预测的图像和真实的图像相比, 误差小的区域变多, 生成质量提高. 模型 4 相对于模型 1 和模型 2, 虽然未在大块黑色区域上有更多的减少, 但是较其他模型结果, 整体的颜色变浅, 如样例 2 的右下角以及样例 3 的中部区域. 模型 4 的设计是针对边缘区域的重建, 边缘区域在图像中大量存在且面积狭小, 体现在图像上即整体的误差颜色变浅.

3.2　Cityscapes 数据集实验

4 个模型在 Cityscapes 数据集上的测试结果如图 8 所示. 分析实验结果可知, Cityscapes 的实验结果和 KITTI 的结果一致. 比如样例 1 中右侧的

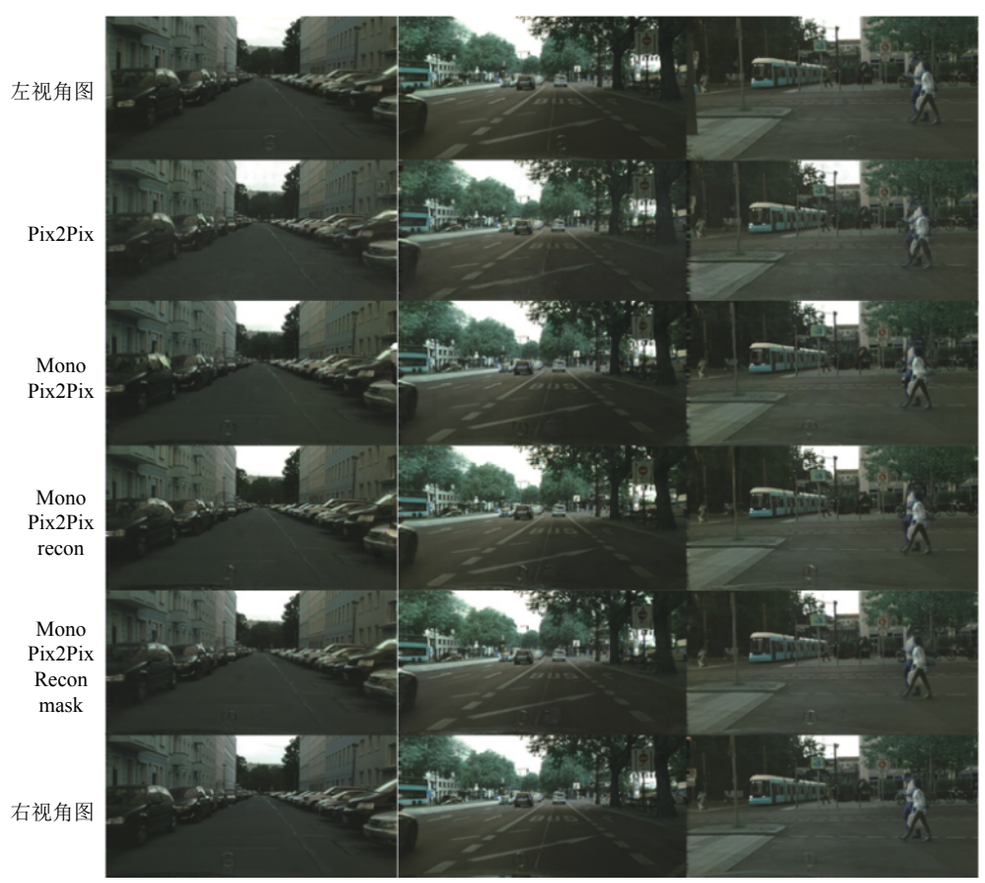


图 8　Cityscapes 数据集下不同模型实验结果对比图

Fig.8　Experiment results of different models on Cityscapes dataset

前车头在 Pix2Pix 模型中生成的曲线不直，2、3、4 模型中明显更接近真实图像；样例 2 中的车道斜线在模型 1 中断断续续且不直，2、3、4 模型中的直线都更直. 总体观察表明，模型 2、3、4 中的生成质量都比模型 1 更高.

和 3.1 节一致，为了比较模型 2、3、4 的生成质量，将生成图像与真实图像的差值处理为灰度图，结果如图 9 所示.

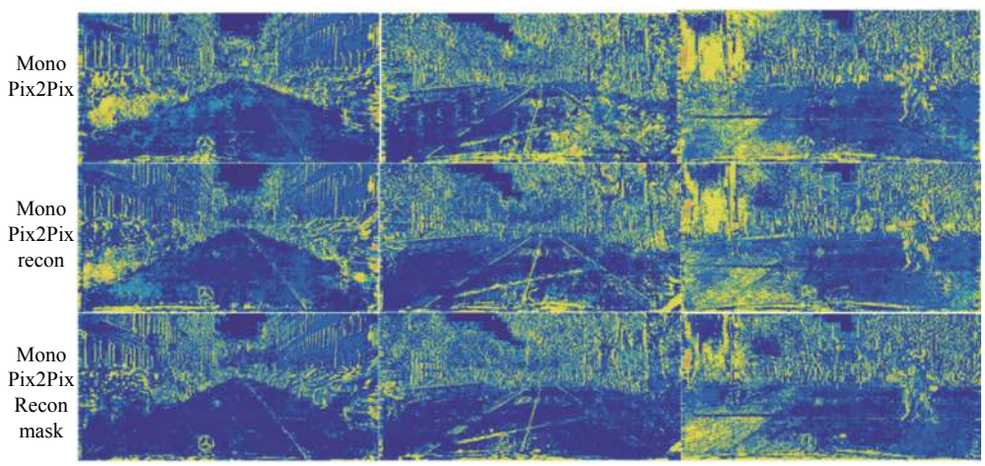


图 9 多种改进模型在 Cityscapes 数据集上的实验对比

Fig.9 Contrasts of experiment results between modified models on Cityscapes dataset

根据 L_1 误差度量的结果图来看，和 KITTI 的实验结果一致. 模型 3 相对于模型 2，黑色区块的面积更大，说明模型 3 的预测结果更接近真实值的区域块更大. 模型 4 在黑色区块的细节上，颜色比模型 3 更黑（比如样例 2 中的路面），说明模型 4 在生成的细节上比模型 3 好.

3.3 综合图表分析

为了方便分析，将对各个数据集、各个参数的实验结果进行定量计算并制作成图表，如图 10 所示.

为了更加量化衡量图像的生成质量好坏，采用衡量图像之间相似度常用的若干指标：均方误差（MSE）、峰值信噪比（PSNR）[25]及结构相似性（SSIM）[23]. MSE 和 PSNR 衡量图像间的像素级相似度，SSIM 衡量图像间的感知相似度. 计算结果如图 10 所示. 图中，recon 为图像补全网络生成的图像，final 版本为根据 STN 得到的中间结果并依据图像补全网络补全的图像. 在 KITTI 的测试中，通过比较性能计算值可以得到以下结论.

1）改进模型 2、3、4 的图像生成质量比基准模型 Pix2Pix 有很大的提高. 在改进的模型中，普遍添加了已经学习好的关于深度信息的先验. 因为网络输入增加了更多的先验信息，图像的生成质量有了很大的提高. 测试结果显示，在添加深度先验信息的实验结果中，MSE 下降了 20%，PSNR 提高了 15.4%，SSIM 提高了 35.5%.

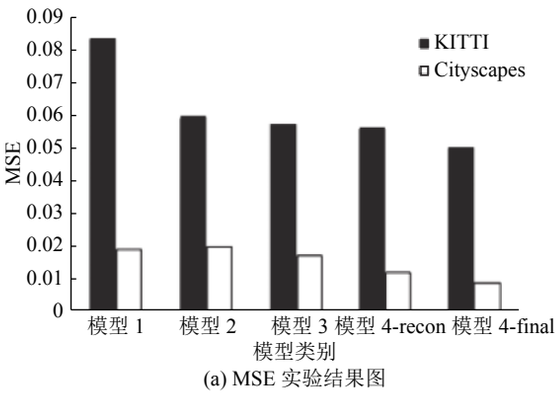
2）模型 2、3、4 之间的性能提升较低. 模型 3 相比较模型 2 在 3 个指标上平均提升约 2%，模型 4 相比较模型 3 提升约 2%. 实验结果表明，重建损失及带有掩码的重建损失对于模型的学习有提升. 先验信息未变，可以利用信息固定导致提升幅度不大.

3）模型 4 的 final 版本测试值有很明显的提高，因为图像中的一部分是从源视角图转换到目标视角来的，保留了源视角图像的高频信息，尤其是在 SSIM 指标上提高较多，因为 SSIM 衡量图像间的感知距离，即高频信息部分.

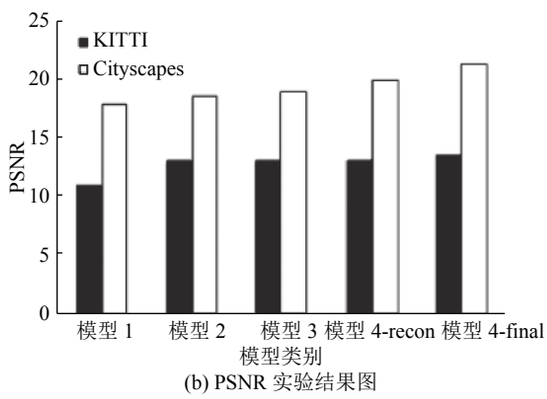
4）模型改进的结果在 2 个实验数据集上都作了测试，2 个数据集在定量指标上表现一致. 横向对比可以发现，模型在 Cityscapes 数据集上的表现优于 KITTI 数据集，尤其是模型 2 到模型 3 的改进. 3 个性能指标在 KITTI 数据集上提高约 3%，在 Cityscapes 数据集上提高 6%~15%. 实验表明，改进模型的性能和测试的数据集类型、训练模型的参数及测试集的随机分割都有一定的关系.

3.4 不同掩码阈值对模型 4 的影响

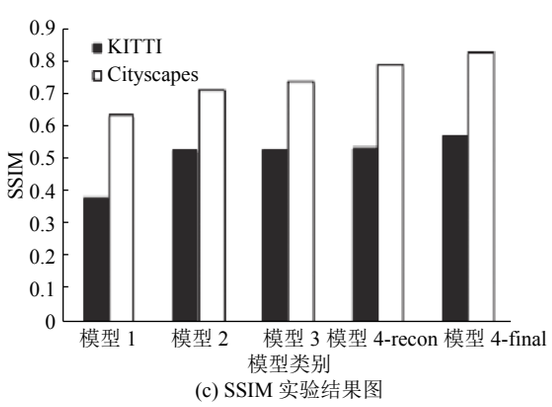
在训练模型的过程中，存在一些超参数的调节. 研究一个重要的参数掩码阈值与模型性能的关系. 掩码阈值用于梯度响应图像二值化的操作，得到图像补全网络需要补全部分的掩码，因此掩码的位置和大小直接影响模型的性能. 由于模型的训练时间非常长，在调节时选用具有代表



(a) MSE 实验结果图



(b) PSNR 实验结果图



(c) SSIM 实验结果图

图 10　不同模型与数据集实验结果定量分析

Fig.10　Quantitative contrasts of experiments of different models and datasets

性的若干值，再根据效果从中选择较好的参数. 得到的掩码图如图 11 所示.

图 11 中的 3 个样例是 3.1 节分析使用的样例. 可以看出，T_m 越小，得到的掩码图的白色区域越多，而白色区域是物体与物体之间的边界处，在视角图像重建的任务中，这些部分是需要重建的部分. 在选取掩码阈值时，考虑图像补全区域的比例. 经过多次试验，首先将参数缩小到合理的 0.001 数量级. 由于设备能力有限，对上述数量级仅测试若干掩码阈值，可得图像重建部分比例及得到的模型 4 的 recon 图像性能图，如图 12 所示. 图中，r 为重建面积比.

通过图 12 可以发现，随着 T_m 的变大，图像补全区域的比例不断下降，图像生成质量先上升后下降. MSE、PNSR 以及 SSIM 的指标在 T_m=0.001 7 时表现最好，在其他 T_m 下的性能低于 T_m=0.001 7 时的结果. 经过分析发现，当 T_m 小时，图像补全区域大于物体边缘区域，所以对超过部分生成图像的损失衡量有误；当 T_m 大时，图像补全区域过小，导致图像补全区域的损失衡量不精确，会出现模型的性能随着 T_m 先上升后降低的情况. 在最终的模型对比中，采用 T_m=0.001 7 时的模型. 通过对比 2 个数据集之间的测试结果发现，T_m 与图像生成质量在 2 个数据集之间是一致的，与 3.3 节的情况类似，KITTI 数据集在测试指标上的响应低于 Cityscapes 的响应.

3.5　恢复图像用于 SLAM 系统的可行性验证

本文的目标是解决在双目系统中，当某一目出现某种干扰因素而不能有效地工作时，通过添加临时替代方案提高系统的容错能力. 将生成的新视角图像代入经典的 ORB-SLAM 系统[4]中，计算相机的轨迹，通过计算与传统双目法求解得到的相机轨迹之间的相对误差，验证了该方法用于容错的可行性. 容错系统采用离线的方式，即先采集好图像，然后离线训练. SLAM 的测试采用离线的方式，即先得到所有的新视角图像，然后进行 SLAM 计算与测试.

对真实的右视角图和生成的右视角图都运行一次 ORB-SLAM 程序，得到每种情况下的相机位姿轨迹图. 计算生成的坐标值与真实的坐标值之间的相对误差，得到的误差如表 1 所示. 表中，x 为轨迹坐标值，Δx 为容错情况与非容错情况下 SLAM 计算的轨迹坐标差值.

根据测试的结果可以发现，通过重建新视角图的方式得到的相机位姿轨迹图的相对误差保持为 10%~20%. 在一定的误差范围内实现了短时间内双目视觉系统在某目失效的情况下，生成新视角缺失图像并重建 SLAM 场景的作用. 在 2 块 GTX 1080Ti GPU 上的实验表明，整个系统每秒可以处理约 5 张图像，系统的实时性需要提高.

4　结　语

本文在经典图像生成网络 Pix2Pix 的基础上，逐步添加了深度估计网络的输入、重建损失以及图像补全损失 3 个方面的改进，得到生成质量越来越好的图像. 首先通过添加深度估计网络和深

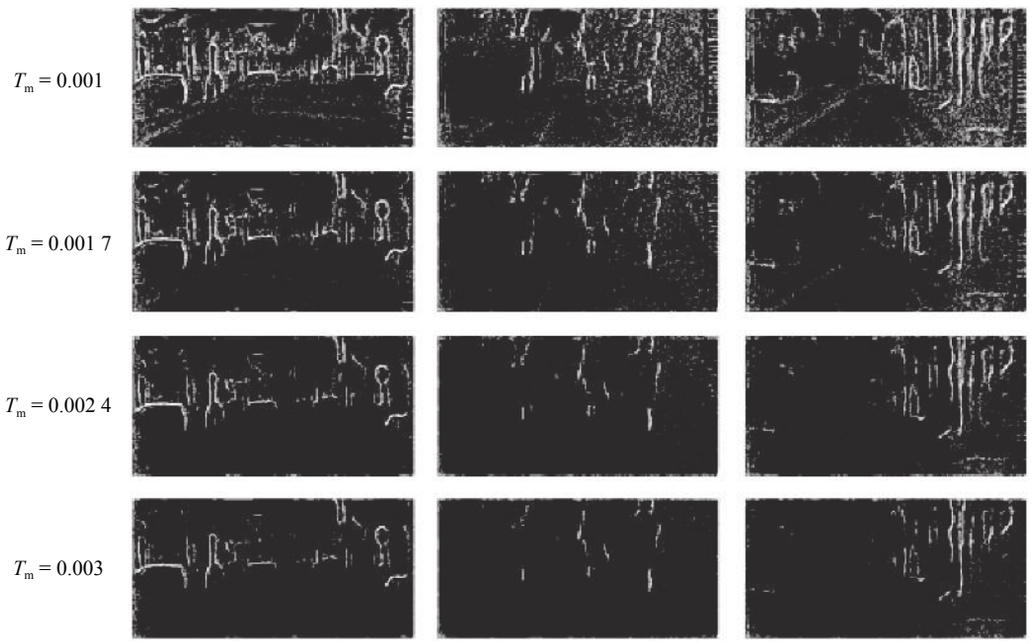


图 11 不同掩码阈值产生的掩码图对比

Fig.11 Mask images generated by different threshold value on depth images

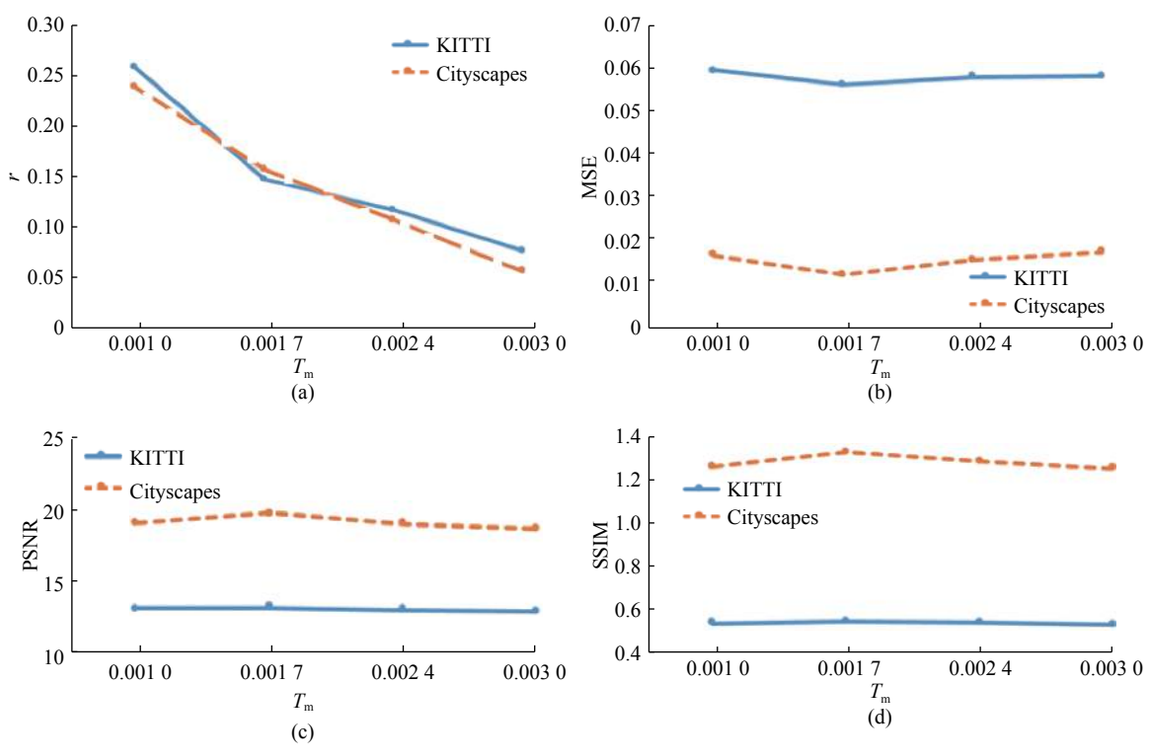


图 12 不同掩码阈值的实验结果对比图

Fig.12 Experimental results of different threshold values

表 1 ORB-SLAM 实验结果

Tab.1 Results of ORB-SLAM experiments

衡量指标	$\lvert\Delta x\rvert/x$	$\sqrt{\Delta x^2}/x$
KITTI	16%	23%
Cityscapes	14%	18%

度信息，得到具有更多先验信息的输入，极大地提高了图像的生成质量；后续添加的基于 STN 网络的图像重建损失，对于 KITTI 数据集和 Cityscapes 数据集的实验结果有了更多的改进；最后加入带有图像补全思想的掩码损失，进一步提高了图像的生成质量. 实验结果在 2 个数据集上表

现一致, 证明了改进结果的有效性. 本文导出深度估计网络的深度信息中间结果, 该相对深度图像可以用于传感器融合, 同时将生成的图像用于双目视觉系统的重建, 证明了图像生成用于 SLAM 容错的可行性和有效性.

参考文献 (References):

[1] SHUM H, KANG S B. Review of image-based rendering techniques [C] // **Visual Communications and Image Processing**. Perth: International Society for Optics and Photonics, 2000, 4067: 2–14.

[2] TATARCHENKO M, DOSOVITSKIY A, BROX T. Multi-view 3D models from single images with a convolutional network [J]. **Knowledge and Information Systems**, 2015, 38(1): 231–257.

[3] DAVISON A J. Real-time simultaneous localisation and mapping with a single camera [C] // **Proceedings Ninth IEEE International Conference on Computer Vision**. Nice: IEEE, 2003: 1403–1410.

[4] MUR-ARTAL R, MONTIEL J M M, TARDOS J D. ORB-SLAM: a versatile and accurate monocular SLAM system [J]. **IEEE Transactions on Robotics**, 2015, 31(5): 1147–1163.

[5] DURRANTWHYTE H F, BAILEY T. Simultaneous localization and mapping [J]. **IEEE Robotics Automat Mag**, 2006, 13(3): 108–117.

[6] LEMAIRE T, BERGER C, JUNG I K, et al. Vision-based SLAM: stereo and monocular approaches [J]. **International Journal of Computer Vision**, 2007, 74(3): 343–364.

[7] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks [C] // **Advances in Neural Information Processing Systems**. Montreal: [s. n.], 2015: 2017–2025.

[8] TATARCHENKO M, DOSOVITSKIY A, BROX T. Single-view to multi-view: reconstructing unseen views with a convolutional network [J]. **Knowledge and Information Systems**, 2015, 38(1): 231–257.

[9] ZHAO B, WU X, CHENG Z Q, et al. Multi-view image generation from a single-view [C]// **Proceedings of the 26th ACM International Conference on Multimedia**. Seoul: ACM, 2018: 383–391.

[10] ZHOU T, TULSIANI S, SUN W, et al. View synthesis by appearance flow [C] // **European Conference on Computer Vision**. Cham: Springer, 2016: 286–301.

[11] PARK E, YANG J, YUMER E, et al. Transformation-grounded image generation network for novel 3d view synthesis [C] // **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Honolulu: IEEE, 2017: 702–711.

[12] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network [C] // **Advances in Neural Information Processing Systems**. Montreal: MIT Press, 2014: 2366–2374.

[13] SHI J, POLLEFEYS M. Pulling things out of perspective [C] // **IEEE Conference on Computer Vision and Pattern Recognition**. Ohio: IEEE, 2014: 89–96.

[14] LIU F, SHEN C, LIN G, et al. Learning depth from single monocular images using deep convolutional neural fields [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2016, 38(10): 2024–2039.

[15] ABRAMS A, HAWLEY C, PLESS R. Heliometric stereo: shape from sun position [C] // **Computer Vision–ECCV 2012.** Berlin: Springer, 2012: 357–370.

[16] FURUKAWA Y, HERNÁNDEZ C. Multi-view stereo: a tutorial [J]. **Foundations and Trends® in Computer Graphics and Vision**, 2015, 9(1/2): 1–148.

[17] RANFTL R, VINEET V, CHEN Q, et al. Dense monocular depth estimation in complex dynamic scenes [C] // **Computer Vision and Pattern Recognition**. Las Vegas: IEEE, 2016.

[18] SCHARSTEIN D, SZELISKI R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms [J]. **International Journal of Computer Vision**, 2002, 47(1-3): 7–42.

[19] WOODHAM R J. Photometric method for determining surface orientation from multiple images [J]. **Optical Engineering**, 1980, 19(1): 1–22.

[20] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency [C] // **Computer Vision and Pattern Recognition**. Honolulu: IEEE, 2017, 2(6): 7.

[21] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks [C] // **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Honolulu: IEEE, 2017: 5967–5976.

[22] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [C] // **International Conference on Medical Image Computing and Computer-Assisted Intervention**. Cham: Springer, 2015: 234–241.

[23] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. **IEEE Transactions on Image Processing**, 2004, 13(4): 600–612.

[24] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: feature learning by inpainting [C] // **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. Las Vegas: IEEE, 2016: 2536–2544.

[25] WANG Y, LI J, LU Y, et al. Image quality evaluation based on image weighted separating block peak signal to noise ratio [C] // **International Conference on Neural Networks and Signal Processing**. Nanjing: IEEE, 2003: 994–997.