工作总结与计划

一、 工作总结（按时间划分）

日期	工作内容			完成情况及主要问题
	上午	下午	晚上	
星期一	选论文	处理杂事	读论文	了解论文要解决的问题，通过描述和插图认识了提出的网络架构及功能
星期二	读论文	读论文	读论文	完成了论文主题方法的阅读，但论文词句较为晦涩，需要再次阅读
星期三	读论文	读论文	读论文	对论文中提到的三维视觉中的算法等进行了了解和学习，并继续阅读论文
星期四	读论文	读论文	休息	大致浏览了论文的附录内容
星期五	读论文	复读论文	写周报	完成论文的阅读 在进行复读后开始写周报
星期六	写周报	读参考文献	学 pytorch	完成周报 学习 pytorch

二、下一步计划（按任务划分）

编号	工作内容	目标	相关配合
1	阅读深度与本文有关的深度估计的论文	12 月 05 日前完成	无
2	准备毕设任务书	12 月 05 日前完成	待定
3			
4			
5			
主要风险	无		

三、个人分析与总结

内容提要	
1	进度方面：本周能有效按照预期计划完成目标
2	课题方面：通过阅读论文，对深度估计有了初步的了解
3	其他思考：在阅读论文外要学习 pytorch，提高英语水平
4	

四、论文总结

论文标题	Normal Assisted Stereo Depth Estimation
作者及单位	Uday Kusupati[1], Shuo Cheng[2], Rui Chen[3], Hao Su[2] 1.The University of Texas at Austin 2.University of California San Diego 3.Tsinghua University
论文出处	2020-CVPR
创新点提炼	本文的创新点可总结如下： 1) 提出了一种新颖的基于 cost-volume 的多视面法线预测网络（Normal Prediction Network, NNet）； 2) 证明了在 cost volume 上对深度估计的管道模型的和法线估计的模型进行共同学习可以促进这两个任务。据我们所知，这是第一项尝试以联合学习的方式解决多视图场景中的深度和法线估计的工作。
个人想法	法线预测对三维重建，深度估计等任务有一定的促进作用，本文侧重探究了深度估计和法线预测联合学习对两个任务的促进作用，对深度估计的探究较少。

论文方法及结论：

1、 论文提出的问题

基于学习的多视图立体方法在有限的视图上已证明具有可竞争的性能。但是，在具有挑战性的情况下，尤其是在构建交叉视图对应关系较为困难时，这些方法仍然无法产生令人满意的结果。

- 多视图立体（Multi-view stereo, MVS）是计算机视觉中最基本的问题之一，基于学习的 MVS 方法已比传统方法（在相对简单的场景下表现得很好）有显著改进。通常这些方法将任务表述为优化问题，**目标是最大程度地减小像素深度差异的总和。但是缺乏几何约束会导致深度预测不准确，尤其是在低纹理或无纹理的区域；**

- 与作为整体几何特性的深度相比，表面法线表示更局部的几何特性，并且可以从视觉外观更容易推断出来。例如，比起绝对深度，人类更容易估计墙壁是否平坦。如图所示，基于学习的 MVS 方法在深度估计上表现不佳，但在法线预测上却表现得更好；

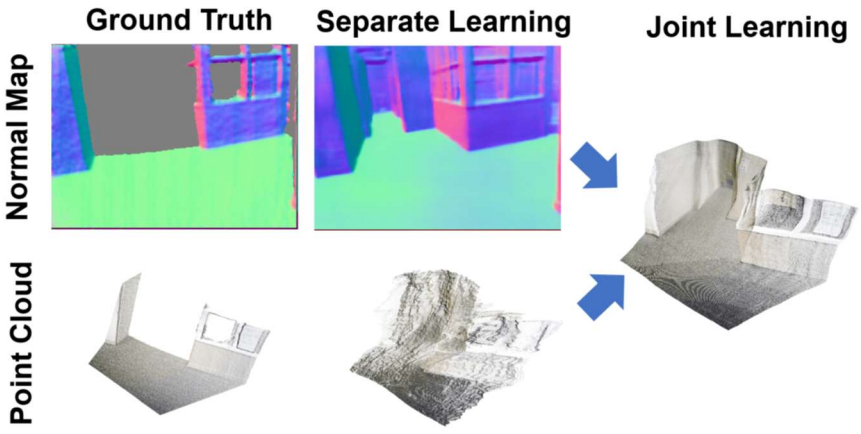

Fig. 1 **Illustration of results of separate learning and joint learning of depth and normal.** While the normal prediction is smooth and accurate, existing state-of-the-art stereo depth prediction result is noisy. Our method improves the prediction quality significantly by joint learning of depth and normal and enforcing consistency.

- 已经有方法尝试将基于法线的几何约束合并到上述优化中以改善单目深度估计（如约束在每个点上预测的法线和从估计深度计算出的切线方向正交，优化深度以与法线保持一致等），但这些方法存在不能估计出最优的深度，或时间耗费长、后期处理失真等问题。

2、 解决的办法

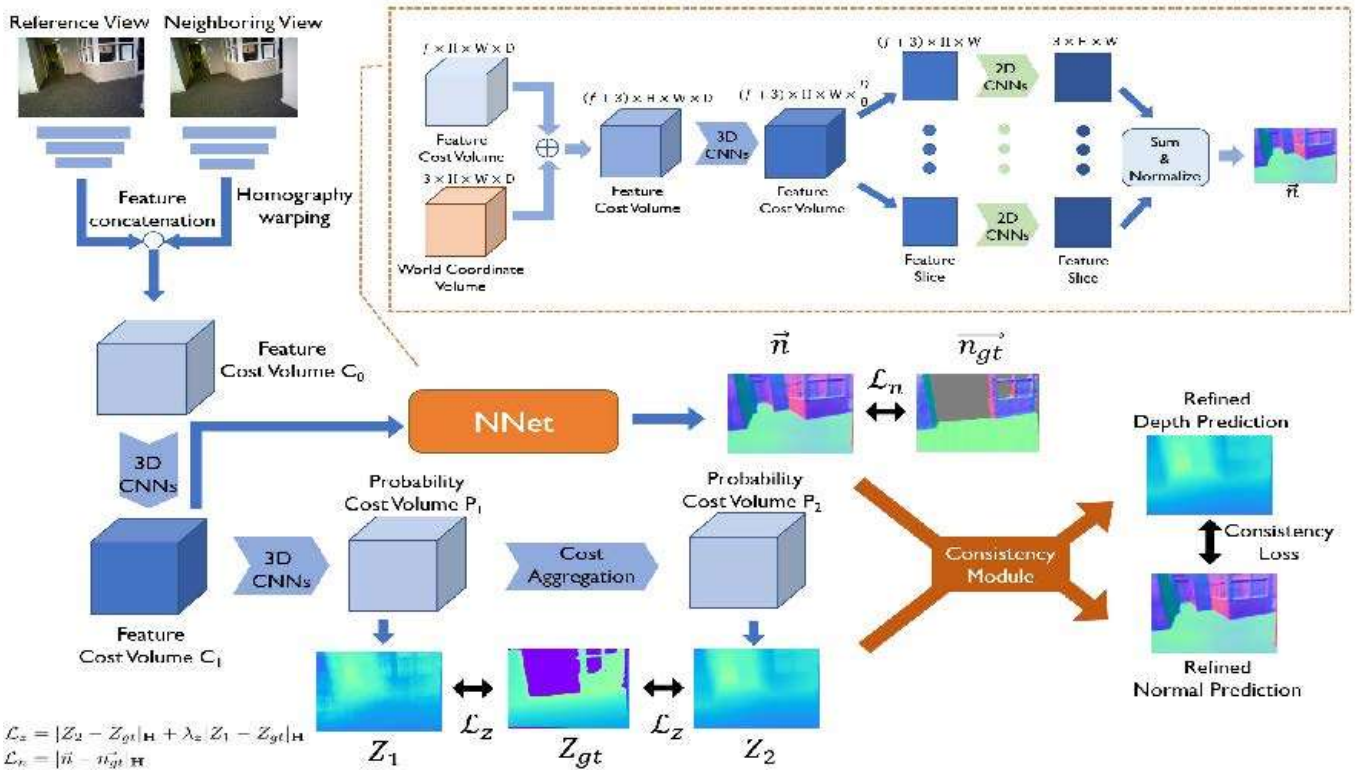

Fig. 2 **Illustration of the pipeline of our method.** We first extract deep image features from viewed images and build a feature cost volume by using feature wrapping. The depth and normal are jointly learned in a supervised fashion. Further we use our proposed consistency module to refine the depth and apply a consistency loss.

本文提出了用于多视图深度和法线估计的端到端的管道模型，如图所示。整个管道模型可以看作两个模块**第一个模块由根据多视图图像特征构建的 cost volume 联合估算深度图和法线图。后续模块通过使用本文提出的一致性损失在预测深度和法线图之间约束一致性来完善深度估计。**在第一个模块中，根据 cost volume 对法线进行联合预测隐式地改善了学习的深度估计模型。第二个模块经过明确训练，可以通过约束一致性来完善估算值。

本文提出一种深度-法线一致性的新公式，以改善训练过程。该一致性是在像素坐标空间中定义的，本文证明了该公式比简单的一致性更好，并且与以前使几何形状一致的方法相比，其性能更好。此约束独立于多视图公式，即使在单视图中，也可用于在任何深度对和法线对上实现一致性。

本文的创新点总结如下：

1) **提出了一种新颖的基于 cost-volume 的多视面法线预测网络（Normal Prediction Network, NNet）**。通过平面扫描构建 3D cost volume，并通过投影将多视图图像信息累积到不同平面上，NNet 可以学习使用正确深度的图像信息准确地推断法线。具有来自多个视图的图像特征的 cost volume 的构建，除了对对应关系以及每个点的深度实施附加约束之外，还包含可用特征的信息。本文提出 cost volume 是一种较好的结构表示形式，有助于更好地学习图像特征以估计基平面的法线。虽然在单幅图像中，网络倾向于过拟合纹理和颜色并显示出较差的泛化性，但由于学习比单幅视图图像更好的抽象性，本文法线估计方法的泛化效果更好；

2) **证明了在 cost volume 上对深度估计的管道模型的和法线估计的模型进行共同学习可以促进这两个任务。**传统的和基于学习的立体方法都受 cost volume 噪声的影响。当基于图像特征的匹配在无纹理的表面上无法提供足够的特征时，该问题尤为明显。本文表明**强制网络从 cost volume 预测准确的法线图会导致 cost volume 表示形式的规范化，从而有助于产生更好的深度估计**。

2.1 Learning based Plane Sweep Stereo

包含多个对象的场景重建需要较大的感受野，网络才能更好地推断上下文信息。由于本文的工作针对场景重建，因而采用最先进的场景重建方法 DPSNet（End-to-end Deep Plane Sweep Stereo）[23] 作为深度估计模块。

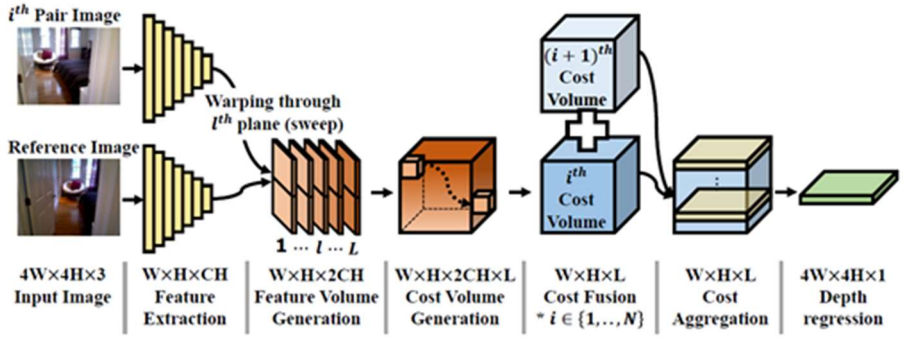

Fig. 3 Overview of the DPSNet pipeline.

网络的输入是同一场景的参考图像 I1 和相邻视图图像 I2，以及相机的内参和外参。首先使用空间金字塔池化模块（聚合多尺度信息）提取深层图像特征。然后通过平面扫描（Plane Sweeping）建立 cost volume，并在其上应用 3D CNN。当存在多个视图时，可以建立多个成本量并将其平均。进一步的上下文感知成本聚合[23]用于正则化携带噪声的 cost volume。最终深度使用 soft argmin [28]从最终的 cost vokume 中回归得到。

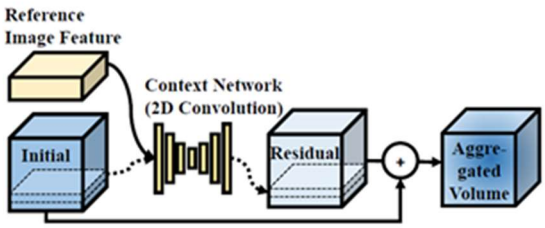

Fig. 4 Illustration of context-aware cost

$$\text{soft argmin} \coloneqq \sum_{d=0}^{D_{max}} d \times \sigma(-c_d)$$

2.2 Cost Volume based Surface Normal Estimation

Cost volume 包含场景中的所有空间信息以及其中的图像特征。Probability volume 对每个像素的候选平面上的深度分布进行建模。在无限的候选平面的极限情况下，精确估计的 probability volume 被证明是基面的隐函数表示，即，取值为 1 表示表面上存在一个点，而其他位置为 0。本文使用使用包含图像级特征的 cost volume C_1 来估计基面的法线图 $\vec{n}$。

给 cost volume C_1，并将每个体素的世界坐标连接到其特征。然后，沿深度维度使用三层 2 步卷积，以将此输入的大小减小为（（f + 3）× H × W × D / 8）并将其称为 C_n。考虑 C_n 中大小为（（f + 3）× H × W）的正平行切片 S_i，将每个切片传入 NNet 。 NNet 包含 7 层 3 × 3 的 2D 卷积层，随着各层加深而使用不同的空洞卷积（1、2、4、6、8、1、1），其感受野也越大。将所有切片的输出相加，并对总和进行归一化以获得法线图的估计。

$$\vec{n} = \frac{\sum_{i=1}^{D/8} NNet(S_i)}{||\sum_{i=1}^{D/8} NNet(S_i)||_2}$$

每个切片包含与在当前切片的感受野中的“幻觉深度”为条件的所有视图中与每个像素的补丁匹配相似度相对应的信息。另外，通过 3D 卷积切片特征会累积一组相邻平面的特征信息。当我们连接世界坐标时，每个平面中每个像素的位置信息都被明确编码为特征。因此，$NNet(S_i)$ 是每个像素的法线估计值，其条件是当前切片的感受野中的深度。对于特定像素，接近真实深度的切片可预测良好的法线估计，而远离真实深度的切片可预测零估计。一种查看方式是，如果来自每个切片的一个像素的法线估计为 $\vec{n}$，则 $\vec{n}$ 的数值大小可以看作是该切片在该像素处的对应概率。 $\vec{n}$ 的方向可以看作是与该切片中像素周围的局部 patch 中的强对应性对齐的向量。

在第一个模块上对 Z_1 和 Z_2 进行真实深度（Z_{gt}）监督训练，对（$\vec{n}$）进行真实法线（$\overrightarrow{n_{gt}}$）监督训练。损失函数（L）定义如下。

$$\mathcal{L}_z = |Z_2 - Z_{gt}|_H + \lambda_z |Z_1 - Z_{gt}|_H$$

$$\mathcal{L}_n = |\vec{n} - \overrightarrow{n_{gt}}|_H$$

$$\mathcal{L} = \mathcal{L}_z + \lambda_n \mathcal{L}_n$$

其中 $|\cdot|H$ 表示 Huber norm。

2.3 Depth Normal Consistency

除了根据 cost volume 共同估算深度和法线之外，本文还使用一种新颖的一致性损失来增强深度估计和法线图之间的一致性。通过利用相机模型、深度图和法线图来估计深度图在像素坐标空间中的空间梯度。我们为 $\left(\frac{\partial Z}{\partial u}, \frac{\partial Z}{\partial v}\right)$ 计算两个估计，并强制它们保持一致。相机模型如下图所示

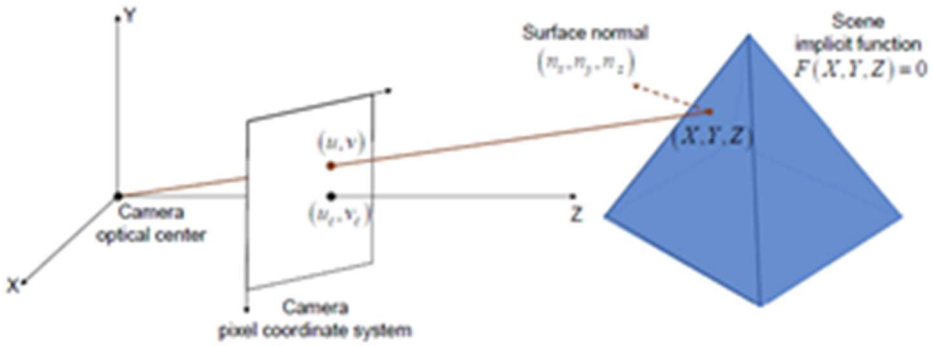

Fig. 5 Camera Model

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} f_x & 0 & u_c \\ 0 & f_y & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

从相机模型可以得出

$$X = \frac{Z(u-u_c)}{f_x} \Longrightarrow \frac{\partial X}{\partial u} = \frac{u-u_c}{f_x}\frac{\partial Z}{\partial u} + \frac{Z}{f_x}$$
$$Y = \frac{Z(v-v_c)}{f_y} \Longrightarrow \frac{\partial Y}{\partial u} = \frac{v-v_c}{f_y}\frac{\partial Z}{\partial x}$$

Estimate 1:

首先可以使用 Sobel 滤波器从深度图计算深度图的空间梯度：

$$\left(\frac{\partial Z}{\partial u},\frac{\partial Z}{\partial v}\right)_1=\left(\frac{\Delta Z}{\Delta u},\frac{\Delta Z}{\Delta v}\right)$$

Estimate 2:

我们假设基面为光滑表面，可以表示为隐函数 F（X; Y; Z）=0。法线图 $\vec{n}$ 是该表面梯度的估计。

$$\vec{n}=(n_x,n_y,n_z)=\left(\frac{\partial F}{\partial X},\frac{\partial F}{\partial Y},\frac{\partial F}{\partial Z}\right)$$
$$\Longrightarrow \frac{\partial Z}{\partial X}=\frac{-n_x}{n_z},\frac{\partial Z}{\partial Y}=\frac{-n_y}{n_z}$$

因此，我们可以通过以下方法得出深度空间梯度的第二个估计值：

$$\left(\frac{\partial Z}{\partial u}\right)_2=\frac{\partial Z}{\partial X}\frac{\partial X}{\partial u}+\frac{\partial Z}{\partial Y}\frac{\partial Y}{\partial u}$$
$$=\frac{\left(\frac{-n_xZ}{n_zf_x}\right)}{1+\left[\frac{n_x(u-u_c)}{n_zf_x}\right]+\left[\frac{n_y(v-v_c)}{n_zf_y}\right]}$$
$$\left(\frac{\partial Z}{\partial v}\right)_2=\frac{\left(\frac{-n_yZ}{n_zf_y}\right)}{1+\left[\frac{n_x(u-u_c)}{n_zf_x}\right]+\left[\frac{n_y(v-v_c)}{n_zf_y}\right]}$$

一致性损失 $\mathcal{L}_c$ 作为两个估计值之间的偏差的 Huber norm 给出

$$\mathcal{L}_c=\left|\left(\frac{\partial Z}{\partial u},\frac{\partial Z}{\partial v}\right)_1-\left(\frac{\partial Z}{\partial u},\frac{\partial Z}{\partial v}\right)_2\right|_H$$

深度空间梯度的第二估计仅取决于所讨论像素的绝对深度，而不取决于相邻像素的深度。从法线图获取局部表面信息，作者认为该信息更准确且更容易估算。该一致性公式不仅在像素附近的相对深度之间施加了约束，而且在绝对深度之间也施加了约束。诸如通过限制从深度图获得的局部表面切线与所估计的法线正交的方法来增强深度图和法线图之间的一致性，这些方法通常会对世界坐标空间中的空间深度梯度施加约束，而本文在像素坐标空间中对它们进行约束。

3、实验结果结论

Method	EPE(↓)	1-pixel error rate(↓)
GCNet	1.80	15.6
PSMNet	1.09	12.1
DPSNet	0.80	8.4
GANet-15	0.84	9.9
GANet-deep	0.78	8.7
GANet-NNet	0.77	8.0
Ours	**0.69**	**7.0**

Table 2. Comparative evaluation of our model on Scene Flow datasets. For all the metrics, lower the better.

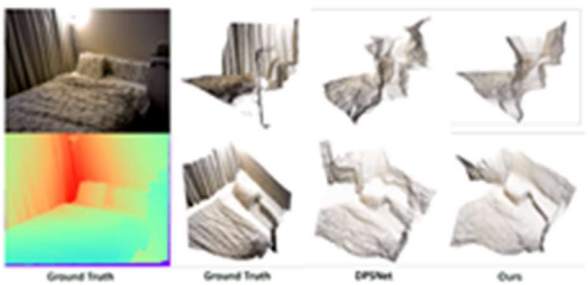

Figure 4. **Visualizing the depths in 3D for SUN3D.** Two views for the point cloud from depth prediction.

Dataset	Method	Abs Rel(↓)	Abs diff(↓)	Sq Rel(↓)	RMSE (↓)	RMSE log(↓)	$\delta < 1.25$ (↑)	$\delta < 1.25^2$ (↑)	$\delta < 1.25^3$ (↑)
MVS (Outdoor)	COLMAP [36]	0.3841	0.8430	1.257	1.4795	0.5001	0.4819	0.6633	0.8401
	DeMoN [40]	0.3105	1.3291	19.970	2.6065	0.2469	0.6411	0.9017	0.9667
	DeepMVS [21]	0.2305	0.6628	0.6151	1.1488	0.3019	0.6737	0.8867	0.9414
	DPSNet-U [23]	0.0813	0.2006	0.0971	0.4419	0.1595	0.8853	0.9454	0.9735
	Ours	**0.0679**	**0.1677**	**0.0555**	**0.3752**	**0.1419**	**0.9054**	**0.9644**	**0.9879**
SUN3D (Indoor)	COLMAP [36]	0.6232	1.3267	3.2359	2.3162	0.6612	0.3266	0.5541	0.7180
	DeMoN [40]	0.2137	2.1477	1.1202	2.4212	0.2060	0.7332	0.9219	0.9626
	DeepMVS [21]	0.2816	0.6040	0.4350	0.9436	0.3633	0.5622	0.7388	0.8951
	DPSNet-U [23]	0.1469	0.3355	0.1165	0.4489	0.1956	0.7812	0.9260	0.9728
	Ours	**0.1271**	**0.2879**	**0.0852**	**0.3775**	**0.1703**	**0.8295**	**0.9437**	**0.9776**
RGBD (Indoor)	COLMAP [36]	0.5389	0.9398	1.7608	1.5051	0.7151	0.2749	0.5001	0.7241
	DeMoN [40]	0.1569	1.3525	0.5238	1.7798	0.2018	0.8011	0.9056	0.9621
	DeepMVS [21]	0.2938	0.6207	0.4297	0.8684	0.3506	0.5493	0.8052	0.9217
	DPSNet-U [23]	0.1508	0.5312	0.2514	0.6952	0.2421	0.8041	0.8948	0.9268
	Ours	**0.1314**	**0.4737**	**0.2126**	**0.6190**	**0.2091**	**0.8565**	**0.9289**	**0.9450**
Scenes11 (Synthetic)	COLMAP [36]	0.6249	2.2409	3.7148	3.6575	0.8680	0.3897	0.5674	0.6716
	DeMoN [40]	0.5560	1.9877	3.4020	2.6034	0.3909	0.4963	0.7258	0.8263
	DeepMVS [21]	0.2100	0.5967	0.3727	0.8909	0.2699	0.6881	0.8940	0.9687
	DPSNet-U [23]	0.0500	0.1515	0.1108	0.4661	0.1164	0.9614	0.9824	0.9880
	Ours	**0.0380**	**0.1130**	**0.0666**	**0.3710**	**0.0946**	**0.9754**	**0.9900**	**0.9947**

Table 1. Comparative evaluation of our model on SUN3D, RGBD, Scenes11 and MVS datasets. For all the metrics except the inlier ratios, lower the better. We use the perfomance of COLMAP, DeMoN, and DeepMVS reported in [23].

Dataset	Method	Abs Rel(↓)	Abs diff(↓)	Sq Rel(↓)	RMSE (↓)
ScanNet	DPSNet	0.1258	0.2145	0.0663	0.3145
	Ours	0.1150	0.2068	0.0577	0.3009
	Ours-$\mathcal{L}_c$	**0.1070**	**0.1946**	**0.0508**	**0.2807**
SUN3D	DPSNet	0.1470	0.3234	0.1071	0.4269
	Ours	0.1332	0.3038	0.0910	0.3994
	Ours-$\mathcal{L}_c$	**0.1247**	**0.2848**	**0.0791**	**0.3671**

Table 3. Comparative evaluation of our consistency loss.

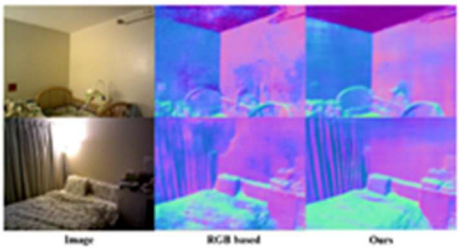

Figure 5. **Surface Normal Estimation.** Test on SUN3D after training on ScanNet. The RGB-based method is from [60]

Method	Mean (↓)	Median (↓)	11.25° (↑)	22.5° (↑)	30° (↑)
RGB-D [53]	14.6	7.5	65.6	81.2	86.2
DC [60]	30.6	20.7	39.2	55.3	60.0
RGB [60]	31.1	**17.2**	**37.7**	58.3	67.1
Ours	**23.1**	18.0	31.1	**61.8**	**73.6**

Table 4. Comparison of normal estimation on ScanNet with single view normal estimation. Note that the RGB-D and depth completion (DC) based methods use ground truth depth. The performances of DC & RGB-D are from [53] and RGB from [60].

Method	Mean (↓)	Median (↓)	11.25° (↑)	22.5° (↑)	30° (↑)
RGB - SUN3D	31.6	25.7	17.9	45.6	57.6
Ours - SUN3D	**22.9**	**17.0**	**34.5**	**63.2**	**73.6**
RGB - MVS	33.3	27.8	11.8	42.4	55.1
Ours - MVS	**27.7**	**22.4**	**23.1**	**52.0**	**63.9**

Table 5. Generalization performance. Both the models were trained on ScanNet (indoor) and tested on SUN3D (indoor) and MVS (outdoor) datasets

4、存在的问题

暂无