

Depth3D: A Model Zoo for Robust Monocular Metric Depth Estimation

1. 方法

为了实现鲁棒单目度量深度估计，我们从现有的 RGB-D 数据集中收集了约 1180 万张图像，并计划向 GitHub 社区发布三个具有竞争力的单目深度估计模型。在本章节中，我们首先介绍本文采用的单目深度估计整体框架及原理，然后介绍本文采用的具体模型结构；最后，我们将提供详细的训练范式设置，包括损失函数、数据集与数据增广。

1.1. 整体框架

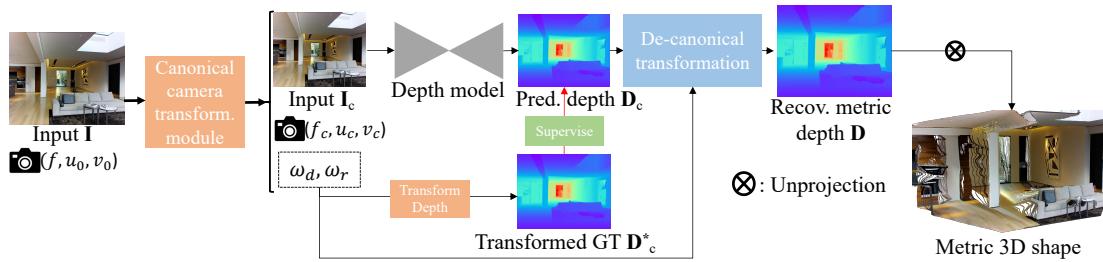


Figure 1. Depth3D 整体框架图。

本文采用的单目深度估计整体框架如 Figure 1 所示。在神经网络方面，本文采用传统 Encoder-Decorder 架构，网络输入为单张 RGB 图片，网络输出该图片对应的深度信息。为了保留单目深度模型的泛化能力，传统方案建模学习仿射不变单目深度，并在实际应用中依赖与稀疏真值深度做最小二乘拟合对齐，回复仿射不变单目深度未知的尺度与偏移量。与传统方案不同，本文通过额外记录拍摄相机内参信息，并通过同时缩放相机焦距和相机拍摄图片深度的原理，将不同相机拍摄的图片，映射至一个统一的标准相机拍摄空间进行深度学习，即：

$$\omega_d = \frac{f_c}{f}, \quad \mathbf{D} = \frac{1}{w_d} \mathbf{D}_c, \quad \mathbf{I}_c = \mathbf{I}$$

在推理阶段，则通过该映射过程的逆过程恢复真实的尺度信息。标准空间的映射原理如 Figure 2 和 Figure 3 所示：

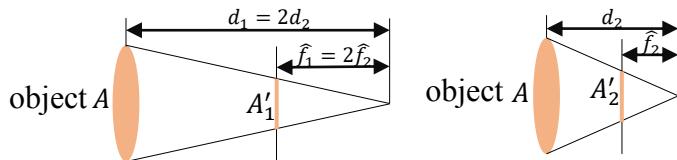


Figure 2 统一缩放相机焦距与深度信息，拍摄图片保持不变。

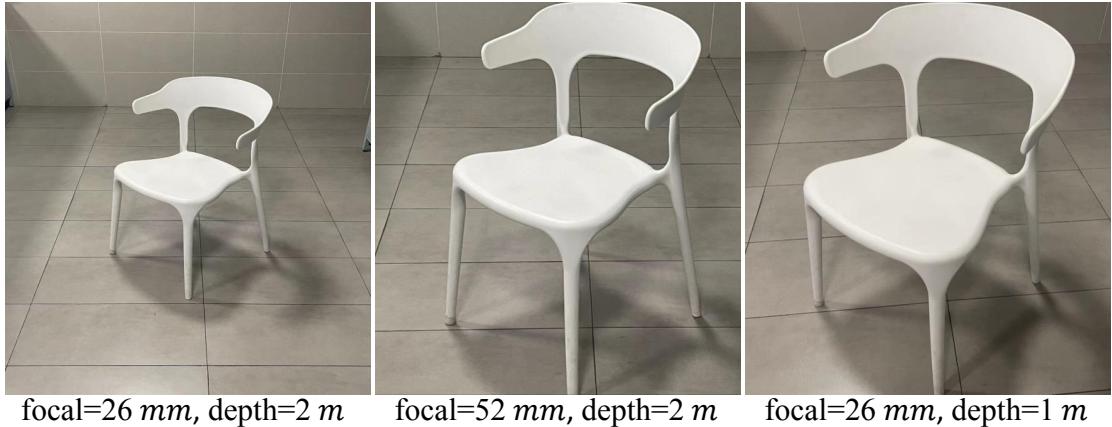


Figure 3. 统一缩放相机内参与深度信息，拍摄图片保持不变。

1.2. 具体模型结构

针对模型结构而言，我们选择了三种模型用于训练鲁棒单目度量深度模型。与 Metric3D 类似，我们采用了具有 ConvNext 骨干网络和 Hourglass 解码器的 UNet 架构。为了实现高质量的深度估计，我们采用了基于 Transformer 的 BEiT 编码器和 DPT 解码器。另外，我们还使用了 Swin2 编码器，以在质量和效率之间取得平衡。

1.2.1. ConvNext

作为一个具有代表性的卷积骨干模型，ConvNext 在包括单目深度估计在内的各种下游任务上取得了令人满意的性能。在 Metric3D 的基础上，我们训练的 ConvNext_{544x1216-L} 模型采用了 ConvNext-large 骨干和一个 Hourglass 解码器，并将 544x1216 大小的图像作为输入。编码器使用在 ImageNet-22K 上预训练的权重进行初始化，每个阶段的通道数和块数分别为(192, 384, 768, 1536)和(3, 3, 27, 3)。对于 Hourglass 解码器，它以通道数为(192, 384, 768, 1536)的特征作为输入，并输出通道数为(128, 128, 256, 512)的特征。为了增强学习到的特征，另外使用了一个由三个下采样分支组成的小型 UNet，每个分支包含两个卷积层和一个转置卷积层。增强后的特征被分为深度特征和预测的置信度图。置信度图定义了深度估计的可信度，深度表示为对数空间中 512 个 bin 的加权和。每个 bin 的概率是在两个额外的卷积层之后通过 softmax 层计算得到的。

1.2.2. BEiT

第二个模型 BEiT_{512x512-L} 负责提供高质量的性能。对于编码器，我们采用 512x512 大小的图像作为输入。与 MiDaS v3.1 类似，我们并非简单地使用 BEiT 的原始视觉 Transformer，而是采用 timm 模型创建函数，并采用和 MiDaS v3.1 相同的 hook 机制。相对于 BEiT 编码器中的 Transformer 块的 absolute hook

position 设置为(5, 11, 17, 23)。连接编码器特征的通道数分别设置为(256, 512, 1024, 1024)，每个阶段的输出通道数为 256。对于特征解码，DPT 解码器的有效性已经得到验证。与原始的 DPT 解码器输出仿射不变的视差不同，我们将其用于预测编码的度量深度的倒数，记为 $disp$ 。为了使训练过程更容易优化，我们将最后输出的 ReLU 函数替换为 ELU 函数，并加上 1 以防止预测值为负。同时我们还使用了一个缩放系数 d_{scale} 和最大深度值 d_{max} 来调整预测深度的范围。预测的编码度量深度与计算得到的度量深度之间的关系如公式所示：

$$depth = \frac{d_{scale}}{disp + d_{scale} / d_{max}}$$

其中， $disp$ 和 $depth$ 分别表示预测的编码度量深度的倒数和计算得到的度量深度，实验中将 d_{scale} 和 d_{max} 设定为 1000 和 300。最终计算得到的度量深度范围为 0 米到 300 米。

1.2.3. Swin2

第三个模型 Swin2_{384x384}-L 的目标是在性能和效率之间取得平衡。对于编码器，我们将 384x384 大小的图像作为输入。Swin2 Transformer 与 BEiT 具有类似的基本实现，但它包含了一个分层编码器。我们选择了 relative hook position 为(1, 1, 17, 1)。解码器与 BEiT 的架构的设置相同。

1.3. 训练范式设置

1.3.1. 损失函数

在训练过程中，我们使用虚拟法线损失 L_{VNL} 、L1 损失 L_1 、天空正则化损失 L_{sky} 、分层深度归一化损失 L_{HDN} 和 L_{HDSN} ，以及边缘的成对法线损失 L_{PWNE} 和平面的成对法线损失 L_{PWNP} ：

$$L_{VNL} = \frac{1}{M_1} \sum_{i=1}^{M_1} |\mathbf{n}_{v,i} - \mathbf{n}_{v,i}^*|, L_1 = \frac{1}{N} \sum_{i=1}^N |d_i - d_i^*|,$$

$$L_{sky} = \frac{1}{N} \sum_{i=1}^N |d_i - d_{sky}^*|,$$

$$L_{HDN} = \frac{1}{M_2^d} \sum_{i=1}^{M_2^d} \left(\frac{1}{|U_i^d|} \sum_{u \in U_i^d} |N_u(d_i) - N_u(d_u^*)| \right),$$

$$L_{HDSN} = \frac{1}{M_2^s} \sum_{i=1}^{M_2^s} \left(\frac{1}{|U_i^s|} \sum_{u \in U_i^s} |N_u(d_i) - N_u(d_u^*)| \right),$$

$$N_u(d_i) = \frac{d_i - \text{median}_u(d)}{\frac{1}{|u|} \sum_{j=1}^{|u|} |d_j - \text{median}_u(d)|},$$

$$L_{\text{PWNE}} = \frac{1}{M_3^{\text{edge}}} \sum_{i=1}^{M_3^{\text{edge}}} |\mathbf{n}_{Ai} \cdot \mathbf{n}_{Bi} - \mathbf{n}_{Ai}^* \cdot \mathbf{n}_{Bi}^*|,$$

$$L_{\text{PWNP}} = \frac{1}{M_3^{\text{plane}}} \sum_{i=1}^{M_3^{\text{plane}}} |\mathbf{n}_{Ai} \cdot \mathbf{n}_{Bi} - \mathbf{n}_{Ai}^* \cdot \mathbf{n}_{Bi}^*|,$$

$$L = \lambda_1 L_{\text{VNL}} + \lambda_2 L_1 + \lambda_3 L_{\text{sky}} + \lambda_4 L_{\text{HDN}} + \lambda_5 L_{\text{HDSN}} + \lambda_6 L_{\text{PWNE}} + \lambda_7 L_{\text{PWNP}}$$

其中，所有带有*的变量表示真实值。 $\mathbf{n}_{v,i}$ 表示采样的虚拟法线， U_i^d 和 U_i^s 分别是根据深度范围和空间域进行采样的位置集合。天空区域的深度值 d_{sky}^* 在ConvNext模型中设置为250米，而在BEiT和Swin2模型中设置为300米。 M_3^{edge} 和 M_3^{plane} 是根据边缘和平面进行采样的。 \mathbf{n}_{Ai} 和 \mathbf{n}_{Bi} 表示采样点对(Ai , Bi)的表面法线。超参数 λ_1 、 λ_2 、 λ_3 、 λ_4 、 λ_5 、 λ_6 、 λ_7 分别设置为0.2、1.0、0.01、2.0、2.0、1.0、1.0。对于部分通过COLMAP等算法获取的或者深度真值尺度未知的数据集，监督时则不采用 L_1 损失函数。

1.3.2. 数据集

为了实现鲁棒的单目度量深度估计，我们需要一个大规模的训练数据集来实现良好的性能和泛化性，并且数据集的多样性应尽可能丰富。因此，我们收集了大约1180万个RGB-D图像对及其相机内参参数，这些数据分布在21个数据集中，如Table 1所示。Quality代表深度真值质量，其中Lidar代表激光雷达等传感器获取的稀疏真值深度，Stereo代表RGB-D等传感器获取的较为稠密的真值深度，SfM代表通过COLMAP等算法获取的尺度未知的深度数据集。我们将部分深度尺度大小不确定的数据集当作SfM数据集处理，即不学习其尺度值，只学习depth分布。部分数据集存储在uint16类型的png格式图片中，Scale代表读取深度值与真实度量深度值之间的尺度值。训练数据包括室内、室外和合成数据，这些数据是通过激光雷达、立体相机和结构光运动算法获取的。在训练过程中，我们根据图像数量和质量将这些数据集分为10组，将每组图像数量重复到相同大小，并从每组中均匀随机采样图像。数据平衡过程非常重要，可以平衡数据集之间的不均匀分布。

Dataset	Images	Scene	Quality	Scale
UASOL	92K	Outdoor	Stereo	300
Cityscapes	109K	Outdoor	Stereo	300

DIML	123K	Outdoor	Stereo	300
KITTI	23K	Outdoor	Lidar	256
Argoverse2	1.5M	Outdoor	Lidar	200
Mapillary	74K	Web Data	SfM	-
Taskonomy	4.0M	Indoor	Stereo	512
Lyft	159K	Outdoor	Lidar	300
DDAD	99K	Outdoor	Lidar	300
Pandaset	49K	Outdoor	Lidar	200
Waymo	1.1M	Outdoor	Lidar	200
DSEC	53K	Outdoor	Lidar	300
DIODE	25K	Indoor/ Outdoor	Stereo	1
Tartanair	613K	Synthetic	Stereo	1
Hypersim	298K	Synthetic	Stereo	1
GraspNet	97K	Indoor	SfM	-
BlendedMVS	133K	Synthetic	Stereo	1
AVD	19K	Indoor	Stereo	1000
NYU	24K	Indoor	Stereo	1000
TUM	8K	Indoor	Stereo	5000
ScanNet	2.5M	Indoor	Stereo	1000
Total	11.8M	-	-	-

Table 1. 训练数据组成

对于测试数据集，我们收集了 8 个高质量的室内和室外真实数据集，如 Table 2 所示。KITTI 和 NuScenes 数据集可以展示在自动驾驶场景中的性能。NYU 包含了室内场景，最大深度约为 10 米。ScanNet 和 7-Scenes 包含了最大深度约为 5 米的小房间。DIODE 和 ETH3D 包含了室内和室外场景，并且深度真值标签的质量很高。iBIMS-1 包含了 100 个不同室内房间的 RGB-D 图像对。由于 ScanNet、7-Scenes 和 NuScenes 包含的图像较多，因此我们分别对其进行了 1000、650 和 1000 张图像的采样用于测试。

Dataset	Images	Scene	Quality	Scale
KITTI	652	Outdoor	Lidar	256
NYU	654	Indoor	Stereo	1000
ScanNet	1000	Indoor	Stereo	1000
7-Scenes	650	Indoor	Stereo	1000
DIODE	771	Indoor/ Outdoor	Stereo	1
ETH3D	454	Indoor/ Outdoor	Stereo	1
iBIMS-1	100	Indoor	Stereo	65535/50
NuScenes	1000	Outdoor	Lidar	200
Total	5281	-	-	-

Table 2. 测试数据组成

1.3.3. 数据增广

在训练过程中，我们使用 *LabelScaleCanonical* 将图像转换到标准空间。它通过同时缩放相机焦距和深度图的比例，将焦距转换为标准值。对于图像外观增强，我们对所有模型使用 *PhotoMetricDistortion*、*Weather* 增强、*RandomBlur* 和 *RGBCompresion*，模拟在不同亮度、对比度、天气条件、运动模糊、图像质量降低等情况下拍摄的图像。对于图像尺寸增强，*ResizeKeepRatio* 在保持纵横比的同时调整图像尺寸，并填充到特定尺寸。*RandomResize* 使用随机采样的比例等比例缩放调整图像大小。*RandomCrop* 从图像中裁剪出指定大小的区域。如果裁剪大小大于图像大小，则使用填充方法保证图片大小。*RandomEdgeMask* 随机屏蔽图像的边缘。*RandomHorizontalFlip* 则根据采样概率水平翻转图像。

对于训练数据，我们依次应用多种增强技术（*LabelScaleCanonical*、*RandomResize*、*RandomCrop*、*RandomEdgeMask*、*RandomHorizontalFlip*、*PhotoMetricDistortion*、*Weather*、*RandomBlur*、*RGBCompresion*）。对于测试数据，我们简单地使用 *LabelScaleCanonical* 将其转换到规范空间，并使用 *ResizeKeepRatio* 进行调整大小。为了获取度量深度估计结果，估计的深度需要转换回原始空间，这是 *LabelScaleCanonical* 的逆过程。

2. 实验

在本章节中，我们首先介绍深度估计评估指标，并在 8 个测试数据集上展示量化指标对比结果。然后，我们针对尺度学习方法、损失函数、训练数据量等多个方面进行消融实验，并在 7-Scenes、ETH3D、iBIMS-1 和 NuScenes 四个数据集上测试性能指标，分析训练有效性。最后，我们提供了 Depth3D 在三个方向上的应用及其可视化结果。

2.1. 深度估计评估指标

对于单目深度估计，我们评估了绝对均相对误差 AbsRel、均方根误差 RMSE 和像素比例 δ_1 指标：

$$\begin{aligned} \text{AbsRel} &= \frac{1}{|D|} \sum_{d \in D} \frac{|d - d^*|}{d^*}, \\ \text{RMSE} &= \sqrt{\frac{1}{|D|} \sum_{d \in D} (d - d^*)^2}, \\ \delta_1 &= \frac{1}{|D|} |\{d \in D | \max\left(\frac{d}{d^*}, \frac{d^*}{d}\right) < 1.25\}| \times 100\% \end{aligned}$$

其中 d 、 d^* 和 D 分别表示预测深度、真实深度和图像的所有预测深度值的集合。排名表示在上述数据集上的平均排名。这些指标将在 8 个测试数据集上进行评估。

2.2. 量化指标性能对比

为了对比量化指标，我们在 DIODE、iBIMS、7-Scenes、ETH3D 和 NuScenes 上评估度量深度估计，与 Adabins、NewCRFs 方法进行了对比。同时，我们在所有测试数据集上测评仿射不变深度的 AbsRel 和 δ_1 性能指标，并与 DPT、MiDaS、LeReS 等知名单目仿射不变深度估计方法进行对比。在测评时，我们提出的方法在测评时不作任何真值深度范围的限制。对于其他方法的性能指标，我们从 Metric3D 论文中获取。

(因表格太宽，Table 8-18 统一存放在文档末尾) 在 Table 8 中，我们与 Adabins、NewCRFs 进行结果对比，并在所有数据集上取得了显著的提升，获得了最佳的效果。在训练过程中，我们发现 *RandomResize* 增强的调整比例会影响深度比例的准确性，并且不同数据集的最佳比例各不相同。因此，在一些数据集（如 7-Scenes 和 iBIMS-1）中带来的提升不如另外三个数据集明显，但仍有不小提升。对于仿射不变深度，我们通过中值匹配（中值）和最小二乘回归（全局）与深度值进行对齐，并将我们的模型与最新的单目深度估计方法进行比较。尽管我们的模型旨在进行度量深度估计，但在仿射不变深度的测评设定下，我们训练的模型超越了 SOTA 方法 DPT 的结果，如 Table 9 所示。值得注意的是，median 中值对齐有时能够比 global 最小二乘对齐取得更好的 AbsRel 性能，例如 NuScenes，这是因为 global 对齐时常常最小化预测深度与真值深度之间的绝对误差，而 AbsRel 指标测评二者的相对误差。

我们训练的三个模型各有优劣。对于 ConvNext_{544x1216}-L 模型，通过使用 *ResizeKeepRatio*，可以兼顾输入具有较大宽高比和正常宽高比的图像，适用于自动驾驶场景。然而，对于室内场景，它通常会在图像两侧填充黑色边框，导致计算成本较高。而对于 BEiT_{512x512}-L 模型，它更适用于室内场景，并承担了高质量性能模型的责任。为了在效果和效率之间取得平衡，使用 Swin2_{384x384}-L 模型输入 384x384 分辨率的图片将是更好的选择。Table 10-18 列出了在八个测试数据集上的相对误差 (AbsRel)、均方根误差 (RMSE) 和 δ_1 的性能指标。值得注意的是，传统的尺度-偏移量恢复方法，即最小二乘回归，通过最小化预测值与真实值之间的绝对深度误差，可能会降低相对性能（例如 AbsRel 和 δ_1 ）。

2.3. 消融实验

为了验证模型各组成部分的有效性，我们在尺度恢复范式、数据量和损失函数三个层面做了消融实验。消融实验均采用两张 4090 训练，每张卡的 batchsize 为 2，所有实验均训练 40 万次迭代。为了平衡数据集之间数据量的不均衡，消融实验和主实验一样将数据集分组，并在调整组间图片至比例相同。所有消融实验均在 7-Scenes、ETH3D、iBIMS-1、NuScenes 四个数据集上测评尺度深度和仿射不变深度的 AbsRel 指标。

2.3.1. 度量深度尺度恢复范式

本文则采用标准空间变换恢复度量深度尺度。为了证明该方法的有效性，本文不采用标准变换的范式做了对比，如 Table 3 所示：

Method	Metric Depth (AbsRel ↓)				Affine-invariant Depth (AbsRel ↓)			
	7Scenes	ETH3D	iBIMS	NuScenes	7Scenes	ETH3D	iBIMS	NuScenes
w/o canonical	0.337	0.485	0.436	0.224	0.086	0.131	0.064	0.179
Full	0.235	0.242	0.162	0.150	0.086	0.107	0.060	0.165

Table 3. 度量深度尺度恢复范式消融实验

标准深度空间的引入，使得网络学习深度尺度准确性极大提升，同时也能在仿射不变深度上带来一定提升。

2.3.2. 数据量

为了验证数据量对性能的影响，本文设置了三种数据集组成，如 Table 4 所示：

Method	Data Component
6 datasets	UASOL, Cityscapes, DIML, KITTI, Argoverse2, Mapillary
13 datasets	UASOL, Cityscapes, DIML, KITTI, Argoverse2, Mapillary, Taskonomy, Lyft, DDAD, Pandaset, Waymo, DSEC, DIODE
Full	UASOL, Cityscapes, DIML, KITTI, Argoverse2, Mapillary, Taskonomy, Lyft, DDAD, Pandaset, Waymo, DSEC, DIODE, Tartanair, Hypersim, GraspNet, BlendedMVS, AVD, NYU, TUM, ScanNet

Table 4. 数据量消融实验设置

值得特别强调的是，本文将数据集分组、调整组间图片至比例相同，并同样训练 40 万次迭代。因此，模型见过的图片数量是相同的，只是多样性和图片所在的域有所区别。如 Table 5 所示，数据多样性的增加会提升深度模型在多数测试集上的性能指标。然而，因为数据组成的变化，可能影响部分数据集上的性能，这种情况在室内数据集和室外数据集的组成不平衡时尤为明显，例如 NuScenes 数据集。

Method	Metric Depth (AbsRel ↓)				Affine-invariant Depth (AbsRel ↓)			
	7Scenes	ETH3D	iBIMS	NuScenes	7Scenes	ETH3D	iBIMS	NuScenes
6 datasets	2.880	0.816	1.966	0.130	0.281	0.160	0.373	0.155
13 datasets	0.375	0.255	0.218	0.146	0.095	0.108	0.065	0.163
Full	0.235	0.242	0.162	0.150	0.086	0.107	0.060	0.165

Table 5. 数据量消融实验

2.3.3. 损失函数

为了验证损失函数对性能的影响，本文设置三种损失函数组成，如 Table 6 所示：

Method	Loss Component
L1	L_1, L_{sky}
L1_norm	$L_1, L_{sky}, L_{VNL}, L_{PWNE}, L_{PWNP}$
Full	$L_1, L_{sky}, L_{VNL}, L_{PWNE}, L_{PWNP}, L_{HDN}, L_{HDSN}$

Table 6. 损失函数消融实验设置

消融实验结果如 Table 7 所示，整体而言，法向量监督 L_{VNL} , L_{PWNE} , L_{PWNP} 和局部归一化监督 L_{HDN} , L_{HDSN} 能够加强结构信息的学习，但有可能影响深度尺度（例如 Metric Depth 栏的 7Scenes 数据集和 NuScenes 数据集）。只使用 L_1 和 L_{sky} 损失函数性能稍低，因为这种范式没有让网络学习真值深度标注的结构信息，只是简单的回归每个 pixel 的深度值。使用深度计算法向量信息，并监督法向量，能够让网络更好的学习图片在三维空间中的空间信息，因此 AbsRel 指标更低。 L_{HDN} 和 L_{HDSN} 能够更好的学习不同深度范围和不同空间范围下的深度分布，因此 Full 能够取得最佳的性能。

Method	Metric Depth (AbsRel ↓)				Affine-invariant Depth (AbsRel ↓)			
	7Scenes	ETH3D	iBIMS	NuScenes	7Scenes	ETH3D	iBIMS	NuScenes
L1	0.232	0.275	0.176	0.156	0.097	0.118	0.079	0.182
L1_norm	0.229	0.263	0.170	0.144	0.090	0.109	0.068	0.171
Full	0.235	0.242	0.162	0.150	0.086	0.107	0.060	0.165

Table 7. 损失函数消融实验

2.4. 应用场景

2.4.1. 度量深度估计

本文在 Table 2 中提及的 8 个测试数据集及网络图片（In the Wild）上测试深度预测的结果，如 Figure 4 所示，本文从单张 RGB 图片预测的深度信息前后关系准确，结构信息与 GT 相近。其中，深度可视化使用 rainbow 色彩，红色代表深度值越大，蓝色代表深度值越小，黑色代表无效值。

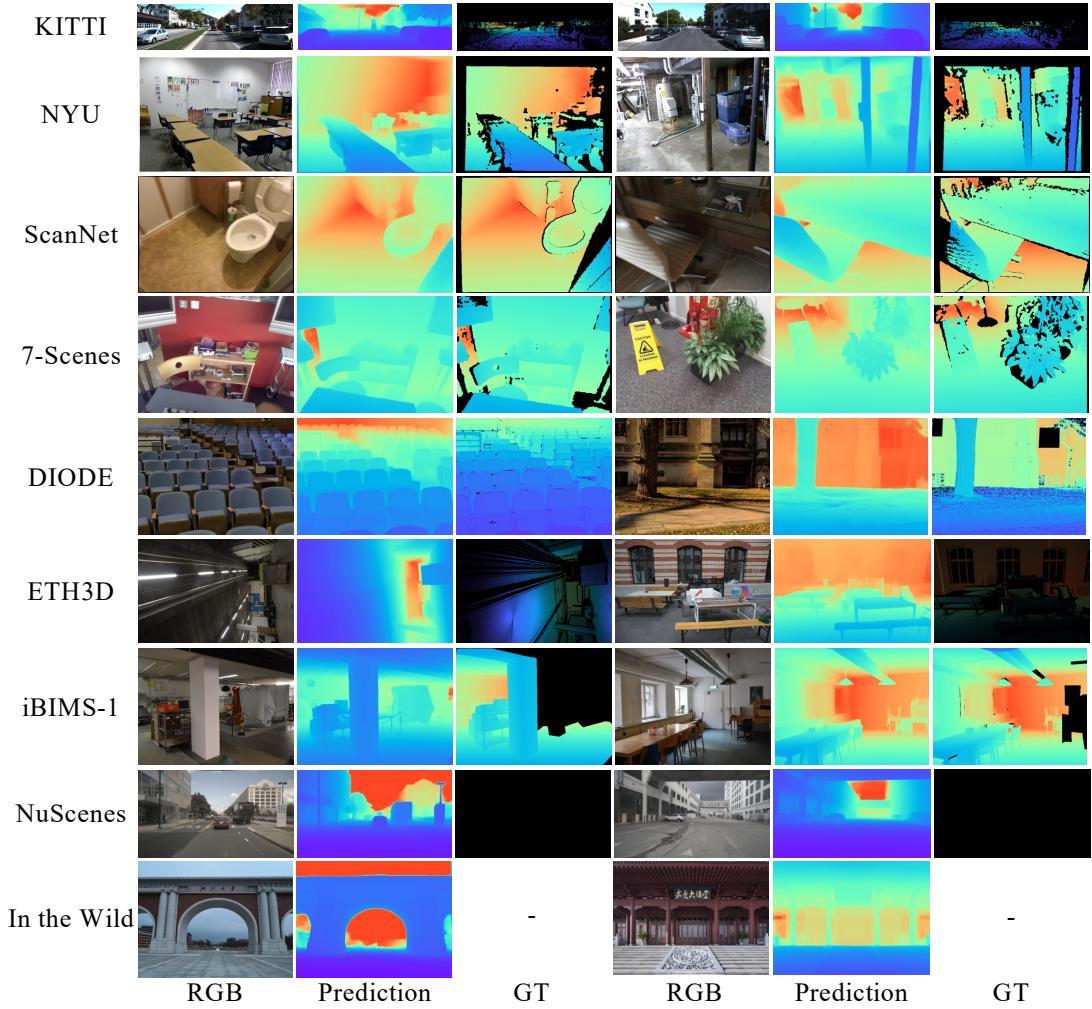


Figure 4. 单目深度估计结果及与其他方法的比较

2.4.2. 图片-点云投影

给定一张 RGB 图片，并输入相机内参参数，本文提出的 Depth3D 方法能够预测该图片的深度信息，并结合相机内参，将 RGB 图片投影为 3D 点云。若图片从网络获取（例如 In the Wild），则统一设置相机内参为 1000，设置光心为图片正中心坐标。Figure 4 中的图片投影得到的点云如 Figure 5 和 Figure 6 所示，本文提出的方法能够鲁棒地从单张图片中恢复出准确的三维点云，其前后关系及几何形状能得到较好的预测。

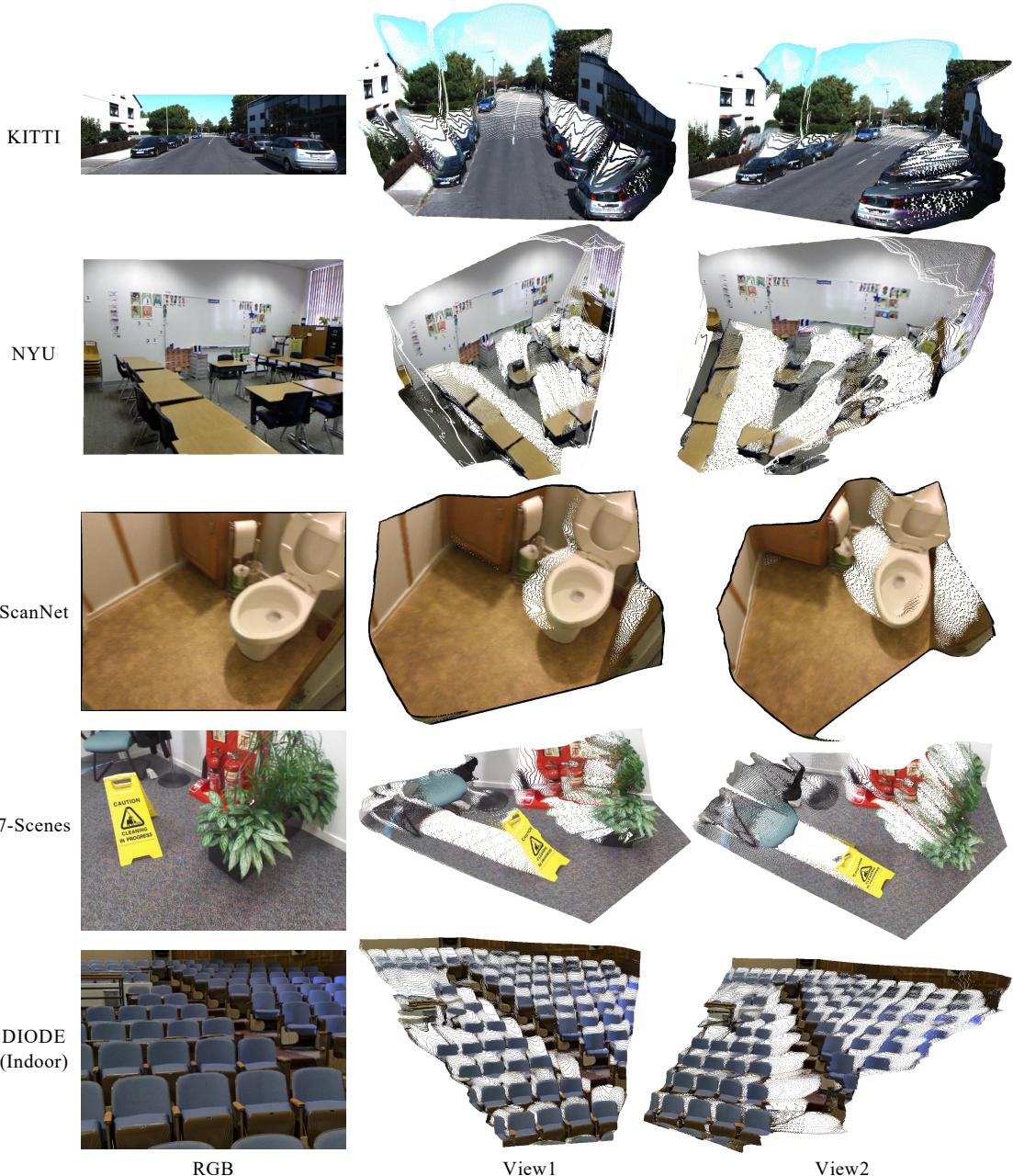


Figure 5. RGB 图片及使用 Depth3D 投影得到的点云

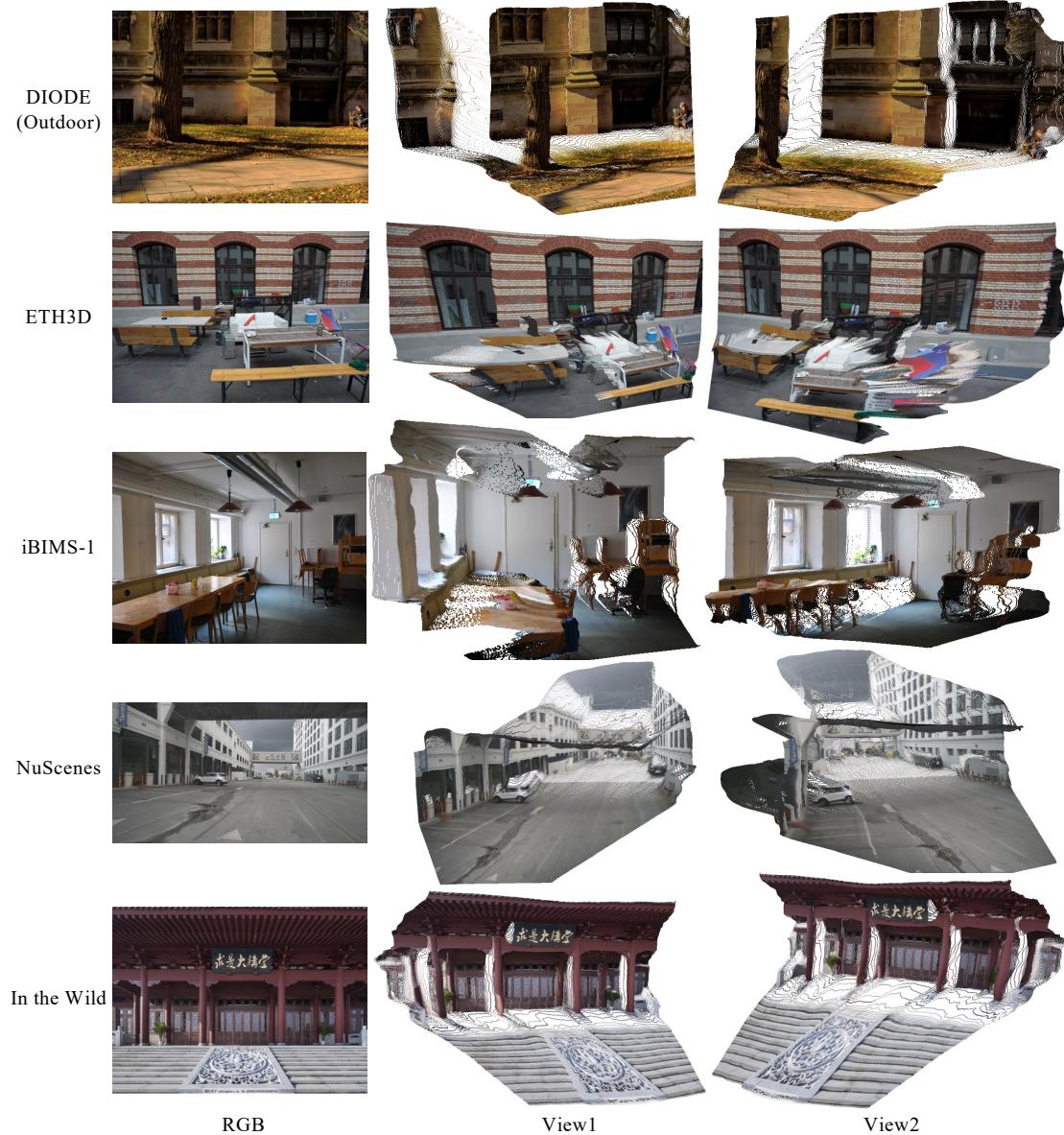


Figure 6. RGB 图片及使用 Depth3D 投影得到的点云

3. 环境安装

3.1. Depth3D 环境安装

Depth3D 安装流程也存放在了 Depth3D 项目代码 README.md 中。

Depth3D 安装环境基于 mmcv 和 pytorch，安装流程如下：

```
conda create -n Depth3D python=3.7
conda activate Depth3D
pip install torch==1.10.0+cu111 torchvision==0.11.0+cu111 -f
https://download.pytorch.org/whl/torch_stable.html
pip install -r requirements.txt
pip install -U openmim
```

```

mim install mmengine
mim install "mmcv-full==1.3.17"
pip install yapf==0.40.1

```

若使用 40 系列显卡，推荐安装 pytorch 2.0 及以上，安装流程如下：

```

conda create -n Depth3D python=3.8
conda activate Depth3D
pip install torch==2.0.0 torchvision=0.15.1
pip install -r requirements.txt
pip install -U openmim
mim install mmengine
mim install "mmcv-full==1.7.1"
pip install yapf==0.40.1

```

3.2. 数据集及其组成

每个数据集由 RGB 图片、Depth 真值深度标注、天空标注文件、annotation 标注文件和 pickle 文件组成。测试数据集共 8 个，包含原始数据及 annotation 标注文件用于存储相机内参及上述各文件路径。我们将提供测试数据集的下载链接，请下载 datasets 文件夹并使用 unzip.sh 脚本解压至 Depth3D 目录下。

(其中 ETH3D 和 DIODE 数据集较大，若下载缓慢，请按照 README_download.md 文件中的引导重构测试数据集。)

3.3. 报告图表复现脚本

在我们提供的代码中，所有用于复现本报告图表指标的脚本均在 Depth3D/scripts/technical_report 目录下。复现教程同时也添加在代码路径 Depth3D/scripts/technical_report/README_reproduction.md 下。

3.3.1. 深度点云可视化复现

在 Depth3D 目录下，运行脚本 bash scripts/technical_report/run_figure4-6.sh 即可推理获取 Figure 4, Figure 5, Figure 6 中的图片。推理完成后，输出文件均保存在 Depth3D/demo_data/outputs_beit 目录下，文件结构如下：

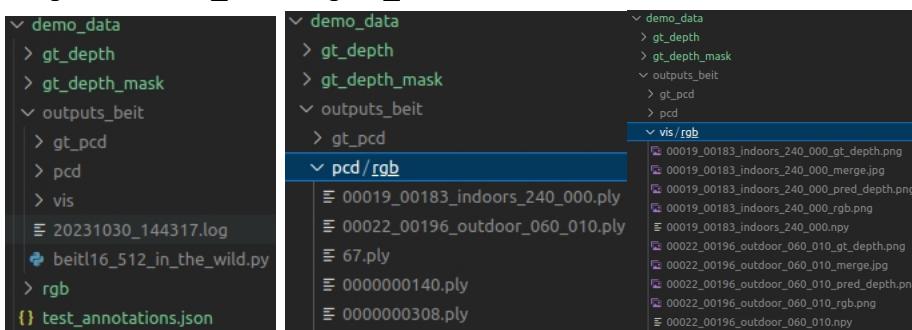


Figure 4 中的 RGB 图片、Prediction Depth 图片、GT Depth 图片均存放在 Depth3D/demo_data/outputs_beit/viz/rgb/ 目录下，Figure 5 和 Figure 6 的结果来自 Depth3D/demo_data/outputs_beit/pcd/rgb/ 目录下的点云使用 MeshLab 软件渲染截图排版获得。

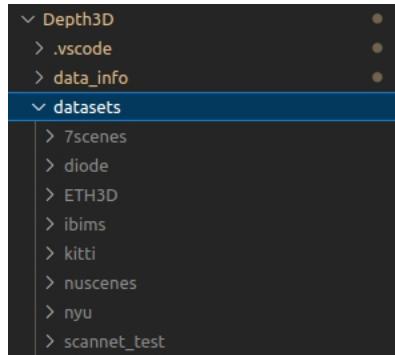
3.3.2. 主实验指标及消融实验指标复现

主实验（Table 8-18）指标及消融实验（Table 3, Table 5, Table 7）指标复现时，在 Depth3D 目录下，运行对应的 run_table*.py 文件，等待数小时即可。例如为获取 Table 3 的结果，运行 python scripts/technical_report/run_table3.py 。运行结果示例如下图所示：

*****Final Evaluation Results of AbsRel are shown as follows*****								
Method	Metric	Metric	Metric	Metric	Affine	Affine	Affine	Affine
	7Scenes	ETH3D	iBIMS	NuScenes	7Scenes	ETH3D	iBIMS	NuScenes
w/o canonical	0.337	0.485	0.436	0.224	0.086	0.131	0.064	0.179
Full	0.235	0.242	0.162	0.15	0.086	0.107	0.06	0.165

Python 脚本的原理为，根据表格所需数据，依次调用 Depth3D/scripts/test 和 Depth3D/scripts/ablation/test 目录下的子实验对应脚本，并将结果存储于 Depth3D/show_dirs/logs_of_technical_report 对应目录下。若检测到之前已运行对应脚本（即监测到对应 log.txt 文件），则跳过此次子实验推理，节省时间。

请注意，请保证 datasets 文件夹按文件结构放置在 Depth3D/datasets 文件夹下，并按下图结构存储。



若需存储在其他路径，则需修改 Depth3D/data_info/public_datasets.py 内各数据集的路径。

3.4. 训练脚本

评测脚本均放在 Depth3D/scripts/train 目录下，训练前请确认 Depth3D/data_info/pretrained_weight.py 文件内的预训练权重路径（代码中默认为 Depth3D/pretrained_weights），其中 ConvNext 模型从 pretrained_weight.py 文件提供的路径中加载在 ImageNet-22K 预训练权重，BEiT 模型和 Swin2 模型在

训练脚本中通过—load-from 选项加载 DPT 预训练权重。若训练数据集没有存放在 Depth3D/datasets 路径下, 请修改 Depth3D/data_info/public_datasets.py 内各数据集的路径及 annotation 文件路径。例如要训练 BEiT 模型, 可在 Depth3D 目录执行 bash scripts/train/train_beit.sh。训练输出 log 示例如下图所示:

```
[10/30 01:59:38] root INFO [Step 188840/400000]
    loss: 2.746,   time: 4.373962,   eta: 14 days, 18:43:42
    decode_EdgetrueNormalLoss: 0.159, decode_HDRandomLoss: 0.282, decode_HDSNRandomLoss: 0.268, decode_L1Loss: 1.71, decode_PMNPlanesLoss: 0.001, decode_SkyRegularizationLoss: 0.032, decode_VNLoss: 0.294
    last_val_errabs_rel: 0.164301, delta: 0.799518, global_abs_rel: 0.204196, global_delta: 0.890341, median_abs_rel: 0.121062, median_delta: 0.880406,
    group0_lr: 0.00000076, group1_lr: 0.00000076, group2_lr: 0.00000076
[10/30 02:00:22] root INFO [Step 188850/400000]
    loss: 3.111,   time: 4.373963,   eta: 14 days, 18:42:58
    decode_EdgetrueNormalLoss: 0.158, decode_HDRandomLoss: 0.318, decode_HDSNRandomLoss: 0.292, decode_L1Loss: 2.010, decode_PMNPlanesLoss: 0.000, decode_SkyRegularizationLoss: 0.029, decode_VNLoss: 0.304
    last_val_errabs_rel: 0.164301, delta: 0.799518, global_abs_rel: 0.204196, global_delta: 0.890341, median_abs_rel: 0.121062, median_delta: 0.880406,
    group0_lr: 0.00000076, group1_lr: 0.00000076, group2_lr: 0.00000076
[10/30 02:01:05] root INFO [Step 188860/400000]
    loss: 3.910,   time: 4.373963,   eta: 14 days, 18:42:14
    decode_EdgetrueNormalLoss: 0.159, decode_HDRandomLoss: 0.286, decode_HDSNRandomLoss: 0.282, decode_L1Loss: 2.848, decode_PMNPlanesLoss: 0.001, decode_SkyRegularizationLoss: 0.043, decode_VNLoss: 0.299
    last_val_errabs_rel: 0.164301, delta: 0.799518, global_abs_rel: 0.204196, global_delta: 0.890341, median_abs_rel: 0.121062, median_delta: 0.880406,
```

消融实验训练代码存放在 Depth3D/scripts/ablation/train/ 目录下, 训练原理及方法与主实验训练脚本相同。

4. 量化性能指标对比表格 (对应 2.2 章节)

Method	DIODE (Indoor)		iBIMS-1		7-Scenes		DIODE (Outdoor)		ETH3D		NuScenes	
	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓
Adabins	0.443	1.963	0.212	0.901	0.218	0.428	0.865	10.35	1.271	6.178	0.445	10.658
NewCRFs	0.404	1.867	0.206	0.861	0.240	0.451	0.854	9.228	0.890	5.011	0.400	12.139
Ours_BEiT	0.183	1.036	0.193	0.675	0.202	0.366	0.331	6.272	0.203	1.425	0.113	6.883
Ours_ConvNext	0.194	0.993	0.228	0.738	0.247	0.412	0.355	6.778	0.237	1.593	0.197	10.940
Ours_Swin2	0.206	1.089	0.226	0.767	0.184	0.351	0.363	6.481	0.246	1.578	0.133	6.636

Table 8. 度量深度估计量化指标对比

Method	KITTI		NYU		ScanNet		DIODE		ETH3D	
	AbsRel ↓	$\delta_1 \uparrow$	AbsRel ↓	$\delta_1 \uparrow$	AbsRel ↓	$\delta_1 \uparrow$	AbsRel ↓	$\delta_1 \uparrow$	AbsRel ↓	$\delta_1 \uparrow$
DiverseDepth	0.190	0.704	0.117	0.875	0.108	0.882	0.376	0.631	0.228	0.694
LeReS	0.149	0.784	0.090	0.916	0.095	0.912	0.271	0.766	0.171	0.777
Omnidata	0.149	0.835	0.074	0.945	0.077	0.935	0.339	0.742	0.166	0.778
HDN	0.115	0.867	0.069	0.948	0.080	0.939	0.246	0.780	0.121	0.833
MiDaS	0.236	0.630	0.111	0.885	0.111	0.886	0.332	0.715	0.184	0.752
DPT-large	0.100	0.901	0.098	0.903	0.078	0.938	0.182	0.758	0.078	0.946
Ours_BEiT (median)	0.067	0.946	0.046	0.976	0.052	0.965	0.198	0.823	0.075	0.951
Ours_ConvNext (median)	0.052	0.972	0.051	0.970	0.063	0.953	0.205	0.815	0.094	0.933
Ours_Swin2 (median)	0.062	0.957	0.049	0.974	0.049	0.967	0.198	0.811	0.085	0.934

Ours_BEiT (global)	0.084	0.933	0.039	0.985	0.042	0.980	0.215	0.823	0.070	0.955
Ours_ConvNext (global)	0.057	0.972	0.043	0.980	0.052	0.970	0.226	0.811	0.090	0.934
Ours_Swin2 (global)	0.065	0.961	0.042	0.981	0.042	0.979	0.214	0.819	0.077	0.947

Table 9. 仿射不变深度估计量化指标对比

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	0.077	0.065	0.099	0.202	0.183	0.331	0.269	0.203	0.193	0.113
Ours_ConvNext	0.055	0.075	0.149	0.246	0.194	0.355	0.287	0.237	0.228	0.197
Ours_Swin2	0.065	0.067	0.093	0.184	0.206	0.363	0.297	0.246	0.226	0.133

Table 10. 度量深度估计（即不做对齐）量化指标 AbsRel

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	3.660	0.245	0.222	0.366	1.036	6.272	4.065	1.425	0.675	6.883
Ours_ConvNext	2.463	0.264	0.297	0.412	0.993	6.778	4.339	1.593	0.738	10.940
Ours_Swin2	2.893	0.251	0.202	0.351	1.089	6.481	4.208	1.578	0.767	6.636

Table 11. 度量深度估计（即不做对齐）量化指标 RMSE

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	0.935	0.971	0.908	0.725	0.680	0.666	0.672	0.699	0.746	0.851
Ours_ConvNext	0.969	0.962	0.827	0.666	0.708	0.678	0.691	0.633	0.673	0.837
Ours_Swin2	0.953	0.969	0.921	0.758	0.692	0.626	0.654	0.591	0.646	0.855

Table 12. 度量深度估计（即不做对齐）量化指标 δ_1

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	0.067	0.046	0.052	0.085	0.068	0.292	0.198	0.075	0.063	0.121
Ours_ConvNext	0.052	0.051	0.063	0.088	0.076	0.299	0.205	0.094	0.074	0.163
Ours_Swin2	0.062	0.049	0.049	0.084	0.072	0.290	0.198	0.085	0.068	0.121

Table 13. 中值对齐单目深度估计量化指标 AbsRel

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	3.468	0.205	0.155	0.232	0.437	5.944	3.652	0.856	0.355	6.505
Ours_ConvNext	2.450	0.217	0.178	0.233	0.523	6.591	4.033	1.140	0.389	10.093
Ours_Swin2	2.863	0.217	0.146	0.238	0.472	6.072	3.711	0.906	0.367	6.632

Table 14. 中值对齐单目深度估计量化指标 RMSE

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	0.946	0.976	0.965	0.925	0.940	0.737	0.823	0.951	0.949	0.867
Ours_ConvNext	0.972	0.970	0.953	0.920	0.923	0.736	0.815	0.933	0.937	0.861
Ours_Swin2	0.957	0.974	0.967	0.928	0.927	0.727	0.811	0.934	0.941	0.866

Table 15. 中值对齐单目深度估计量化指标 δ_1

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	0.084	0.039	0.042	0.079	0.059	0.329	0.215	0.070	0.052	0.142
Ours_ConvNext	0.057	0.043	0.052	0.082	0.069	0.340	0.226	0.090	0.059	0.229
Ours_Swin2	0.065	0.042	0.042	0.078	0.062	0.325	0.214	0.077	0.055	0.158

Table 16. 最小二乘对齐单目深度估计量化指标 AbsRel

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	2.983	0.173	0.128	0.208	0.353	5.354	3.246	0.669	0.295	5.831
Ours_ConvNext	2.225	0.183	0.145	0.210	0.378	5.673	3.441	0.839	0.313	6.857
Ours_Swin2	2.612	0.182	0.125	0.205	0.369	5.347	3.249	0.714	0.297	5.973

Table 17. 最小二乘对齐单目深度估计量化指标RMSE

Method	KITTI	NYU	ScanNet	7-Scenes	DIODE (indoor)	DIODE (outdoor)	DIODE	ETH3D	iBIMS-1	NuScenes
Ours_BEiT	0.933	0.985	0.980	0.935	0.954	0.728	0.823	0.955	0.969	0.835
Ours_ConvNext	0.972	0.980	0.970	0.929	0.942	0.716	0.811	0.934	0.962	0.708
Ours_Swin2	0.961	0.981	0.979	0.935	0.951	0.722	0.819	0.947	0.967	0.803

Table 18. 最小二乘对齐单目深度估计量化指标 δ_1