Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation

Metric3D v2: 用于零样本度量深度和表面法线估计的通用单目几何基础模型

Mu Hu [1*], Wei Yin [2*†], Chi Zhang [3], Zhipeng Cai [4], Xiaoxiao Long [5‡], Hao Chen [6], Kaixuan Wang [1], Gang Yu [7], Chunhua Shen [6], Shaojie Shen [1]

胡牧 [1*]，尹伟 [2*†]，张弛 [3]，蔡志鹏 [4]，龙笑笑 [5‡]，陈浩 [6]，王凯旋 [1]，余刚 [7]，沈春华 [6]，沈少杰 [1]

Abstract-We introduce Metric3D v2, a geometric foundation model for zero-shot metric depth and surface normal estimation from a single image, which is crucial for metric 3D recovery. While depth and normal are geometrically related and highly complimentary, they present distinct challenges. State-of-the-art (SoTA) monocular depth methods achieve zero-shot generalization by learning affine-invariant depths, which cannot recover real-world metrics. Meanwhile, SoTA normal estimation methods have limited zero-shot performance due to the lack of large-scale labeled data. To tackle these issues, we propose solutions for both metric depth estimation and surface normal estimation. For metric depth estimation, we show that the key to a zero-shot single-view model lies in resolving the metric ambiguity from various camera models and large-scale data training. We propose a canonical camera space transformation module, which explicitly addresses the ambiguity problem and can be effortlessly plugged into existing monocular models. For surface normal estimation, we propose a joint depth-normal optimization module to distill diverse data knowledge from metric depth, enabling normal estimators to learn beyond normal labels. Equipped with these modules, our depth-normal models can be stably trained with over 16 million of images from thousands of camera models with different-type annotations, resulting in zero-shot generalization to in-the-wild images with unseen camera settings. Our method currently ranks the 1st on various zero-shot and non-zero-shot benchmarks for metric depth, affine-invariant-depth as well as surface-normal prediction, shown in Fig. 1. Notably, we surpassed the ultra-recent MarigoldDepth and DepthAnything on various depth benchmarks including NYUv2 and KITTI. Our method enables the accurate recovery of metric 3D structures on randomly collected internet images, paving the way for plausible single-image metrology. The potential benefits extend to downstream tasks, which can be significantly improved by simply plugging in our model. For example, our model relieves the scale drift issues of monocular-SLAM (Fig. 3), leading to high-quality metric scale dense mapping. These applications highlight the versatility of Metric3D v2 models as geometric foundation models. Our project page is at https://JUGGHM.github.io/Metric3Dv2.

摘要—我们推出了 Metric3D v2，这是一种用于从单张图像进行零样本度量深度和表面法线估计的几何基础模型，这对于度量三维重建至关重要。虽然深度和法线在几何上相关且高度互补，但它们带来了不同的挑战。目前最先进 (SoTA) 的单目深度方法通过学习仿射不变深度实现零样本泛化，但无法恢复真实世界的度量。同时，由于缺乏大规模标注数据，最先进的法线估计方法的零样本性能有限。为解决这些问题，我们针对度量深度估计和表面法线估计提出了解决方案。对于度量深度估计，我们表明零样本单视图模型的关键在于解决来自各种相机模型和大规模数据训练的度量模糊性。我们提出了一个规范相机空间变换模块，该模块明确解决了模糊性问题，并且可以轻松插入现有的单目模型中。对于表面法线估计，我们提出了一个联合深度 - 法线优化模块，以从度量深度中提取多样化的数据知识，使法线估计器能够超越法线标签进行学习。配备这些模块后，我们的深度 - 法线模型可以使用来自数千个具有不同类型注释的相机模型的超过 1600 万张图像进行稳定训练，从而实现对具有未见相机设置的野外图像的零样本泛化。如图 1 所示，我们的方法目前在各种零样本和非零样本基准测试中，在度量深度、仿射不变深度以及表面法线预测方面均排名第一。值得注意的是，在包括 NYUv2 和 KITTI 在内的各种深度基准测试中，我们超越了最近的 MarigoldDepth 和 DepthAnything。我们的方法能够在随机收集的互联网图像上准确恢复度量三维结构，为可行的单图像计量学铺平了道路。其潜在好处延伸到下游任务，只需插入我们的模型即可显著改善这些任务。例如，我们的模型缓解了单目 SLAM 的尺度漂移问题 (图 3)，实现了高质量的度量尺度密集映射。这些应用凸显了 Metric3D v2 模型作为几何基础模型的通用性。我们的项目页面为 https://JUGGHM.github.io/Metric3Dv2。

Index Terms-Monocular metric depth estimation, surface normal estimation, 3D scene shape estimation

关键词—单目度量深度估计，表面法线估计，三维场景形状估计

1 INTRODUCTION

1 引言

Monocular metric depth and surface normal estimation is the task of predicting absolute distance and surface direction from a single image. As crucial 3D representations, depth and normals are geometrically related and highly complementary. While metric depth excels in capturing data at scale, surface normals offer superior preservation of local geometry and are devoid of metric ambiguity compared to metric depth. These unique attributes render both depth and surface normals indispensable in various computer vision applications, including 3D reconstruction [1], [2], [3], neural rendering (NeRF) [4], [5], [6], [7], autonomous driving [8], [9], [10], and robotics [11], [12], [13]. Currently, the community still lacks a robust, generalizable geometry foundation model [14], [15], [16] capable of producing high-quality metric depth and surface normal from a single image.

单目度量深度和表面法线估计是从单张图像预测绝对距离和表面方向的任务。作为至关重要的三维表示，深度和法线在几何上相关且高度互补。虽然度量深度在捕捉大规模数据方面表现出色，但表面法线在保留局部几何信息方面更胜一筹，并且与度量深度相比没有度量模糊性。这些独特的属性使得深度和表面法线在各种计算机视觉应用中不可或缺，包括三维重建 [1]、[2]、[3]，神经渲染 (NeRF) [4]、[5]、[6]、[7]，自动驾驶 [8]、[9]、[10] 以及机器人技术 [11]、[12]、[13]。目前，该领域仍然缺乏一种强大的、可泛化的几何基础模型 [14]、[15]、[16]，能够从单张图像中生成高质量的度量深度和表面法线。

Metric depth estimation and surface normal estimation confront distinct challenges. Existing depth estimation methods are categorized into learning metric depth [17], [18], [19], [20], relative depth [21], [22], [23], [24], and affine-invariant depth [25], [26], [27], [28], [29]. Although the metric depth methods [17], [18], [19], [20], [30] have achieved impressive accuracy on various benchmarks, they must train and test on the dataset with the same camera intrinsics. Therefore, the training datasets of metric depth methods are often small, as it is hard to collect a large dataset covering diverse scenes using one identical camera. The consequence is that all these models generalize poorly in zero-shot testing, not to mention the camera parameters of test images can vary too. A compromise is to learn the relative depth [21], [23], which only represents one point being further or closer to another one. The application of relative depth is very limited. Learning affine-invariant depth finds a trade-off between the above two categories of methods, i.e. the depth is up to an unknown scale and shift. With large-scale data, they decouple the metric information during training and achieve impressive robustness and generalization ability, such as MiDaS [27], DPT [28], LeReS [25], [26], HDN [29]. The problem is the unknown shift will cause 3D reconstruction distortions [26] and non-metric depth cannot satisfy various downstream applications.

度量深度估计和表面法线估计面临着不同的挑战。现有的深度估计方法可分为学习度量深度 [17]、[18]、[19]、[20]、相对深度 [21]、[22]、[23]、[24] 以及仿射不变深度 [25]、[26]、[27]、[28]、[29]。尽管度量深度方法 [17]、[18]、[19]、[20]、[30] 在各种基准测试中取得了令人瞩目的精度，但它们必须在具有相同相机内参的数据集上进行训练和测试。因此，度量深度方法的训练数据集通常较小，因为很难使用同一台相机收集涵盖不同场景的大型数据集。结果是，所有这些模型在零样本测试中的泛化能力较差，更不用说测试图像的相机参数也可能会有所不同。一种折中的方法是学习相对深度 [21]、[23]，它仅表示一个点相对于另一个点是更远还是更近。相对深度的应用非常有限。学习仿射不变深度在上述两类方法之间找到了一个平衡点，即深度存在一个未知的尺度和偏移。利用大规模数据，它们在训练过程中解耦了度量信息，并实现了令人印象深刻的鲁棒性和泛化能力，例如 MiDaS[27]、DPT[28]、LeReS[25]、[26]、HDN[29]。问题在于未知的偏移会导致三维重建失真 [26]，并且非度量深度无法满足各种下游应用的需求。

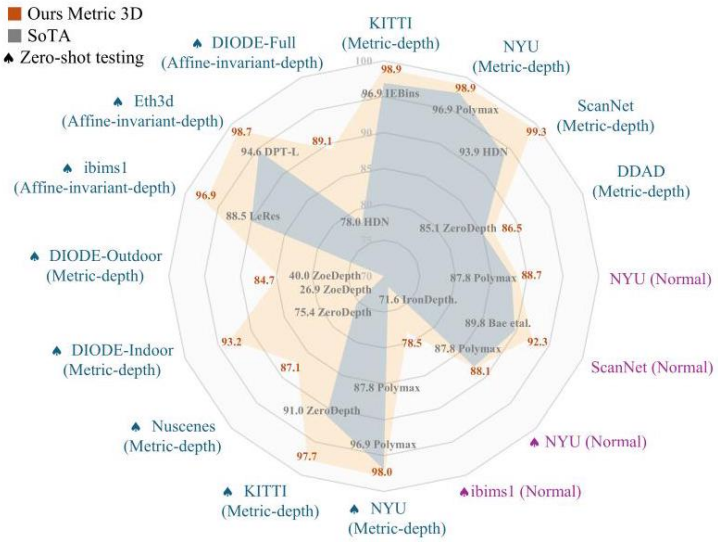

Fig. 1 - Comparisons with SoTA methods on 16 depth and normal benchmarks. Radar-map of our Metric3D V2 v.s. SoTA methods from different works, on (1) Metric depth benchmarks, see '(Metric-depth)'. (2) Affine-invariant depth benchmarks, see '(Affine-invariant-depth)'. (3) Surface normal benchmarks, see '(Normal)'. Zero-shot testing is denoted by ' Φ ' . Here δ_1 percentage accuracy is used for depth benchmarks and 30° percentage

accuracy is for normal. Both higher values are for better performance. We establish new SoTA on a wide range of depth and normal benchmarks.

图 1 - 在 16 个深度和法线基准测试中与最先进方法的比较。我们的 Metric3D V2 与来自不同工作的最先进方法在 (1) 度量深度基准测试 (见‘(Metric-depth)’)、(2) 仿射不变深度基准测试 (见‘(Affine-invariant-depth)’) 和 (3) 表面法线基准测试 (见‘(Normal)’) 上的雷达图。零样本测试用’Φ’表示。这里深度基准测试使用 δ_1 百分比精度，法线使用 30° 百分比精度。数值越高表示性能越好。我们在广泛的深度和法线基准测试中建立了新的最先进水平。

- * Equal contribution.
- * 同等贡献。
- †WY is the project lead (yvanwy@outlook.com).
- †WY 是项目负责人 (yvanwy@outlook.com)。
- ‡XL is the corresponding author (xxlong@connect.hku.hk).
- ‡XL 是通讯作者 (xxlong@connect.hku.hk)。
- Contact MH for technical concerns (mhuam@connect.ust.hk).
- 如有技术问题请联系 MH(mhuam@connect.ust.hk)。
- [1] HKUST [2] Adelaide University
- [1] 香港科技大学 [2] 阿德莱德大学
- [3] Westlake University [4] Intel [5] HKU [6] Zhejiang University [7] Tencent

Submitted for review on Feb. 29th, 2024.
- [3] 西湖大学 [4] 英特尔 [5] 香港大学 [6] 浙江大学 [7] 腾讯于 2024 年 2 月 29 日提交审核。

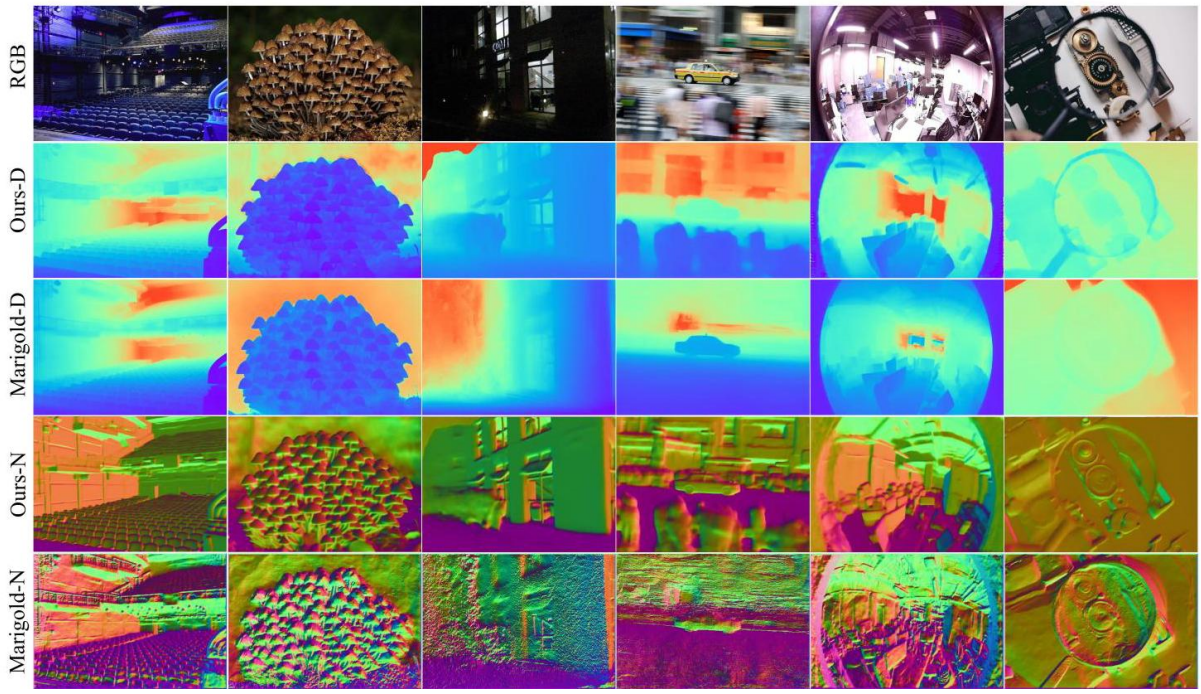

Fig. 2 - Surface normal (N) and monocular depth (D) comparisons on diverse web images. Our method, directly estimating metric depths and surface normals, shows powerful generalization in a variety of scenarios, including indoor, outdoor, poor-visibility, motion blurred, and fisheye images. Visualized results come from our ViT-large-backbone estimator. Marigold is a strong and robust diffusion-based monocular depth estimation method, but its recovered surface normals from the depth show various artifacts.

图 2 - 在不同网络图像上的表面法线 (N) 和单目深度 (D) 比较。我们的方法直接估计度量深度和表面法线，在各种场景中表现出强大的泛化能力，包括室内、室外、低能见度、运动模糊和鱼眼图像。可视化结果来自我们基于 ViT 大骨干网络的估计器。Marigold 是一种强大且鲁棒的基于扩散的单目深度估计方法，但它从深度恢复的表面法线存在各种伪影。

In the meantime, these models cannot generate surface normals. Although lifting depths to 3D point clouds can do so, it places high demands on the accuracy and fine details of predicted depths. Otherwise, various artifacts will remain in such transformed normals. For example, Fig. 2 shows noisy normals from Marigold [31] depths, which excels in producing high-resolution fine depths. Instead of direct transformation, state-of-the-art (SoTA) surface normal estimation methods [32], [33], [34] tend to train estimators on high-quality normal annotations. These annotations, unlike sensor-captured ground-truth (GT), are derived from meticulously and densely reconstructed scenes, which have extremely rigorous requirements for both the capturing equipment and the scene. Consequently, data sources primarily consist of either synthetic creation or 3D indoor reconstruction [35]. Real and diverse outdoor scenes are exceedingly rare. (refer to our data statistics in Tab. 5). Limited by this label deficiency, SoTA surface normal methods [32], [33], [34] typically struggle with strong zero-shot generalization. This work endeavors to tackle these challenges by developing a multi-task foundation model for zero-shot, single view, metric depth, and surface normal estimation.

与此同时，这些模型无法生成表面法线。虽然将深度提升为 3D 点云可以做到这一点，但这对预测深度的精度和精细细节提出了很高的要求。否则，在这种转换后的法线中会残留各种伪影。例如，图 2 显示了 Marigold[31] 深度生成的有噪声的法线，而 Marigold 在生成高分辨率精细深度方面表现出色。与直接转换不同，最先进的 (SoTA) 表面法线估计方法 [32]、[33]、[34] 倾向于在高质量的法线标注上训练估计器。这些标注与传感器捕获的真实值 (GT) 不同，它们来自精心且密集重建的场景，这对捕获设备和场景都有极其严格的要求。因此，数据源主要包括合成创建或 3D 室内重建 [35]。真实且多样的室外场景极为罕见 (请参考表 5 中的数据统计)。受限于这种标注不足，最先进的表面法线方法 [32]、[33]、[34] 通常难以实现强大的零样本泛化能力。本工作致力于通过开发一种用于零样本、单视图、度量深度和表面法线估计的多任务基础模型来解决这些挑战。

We propose targeted solutions for the challenges of zero-shot metric depth and surface normal estimation. For metric-scale recovery, we first analyze the metric ambiguity issues in monocular depth estimation and study different camera parameters in depth, including the pixel size, focal length, and sensor size. We observe that the focal length is the critical factor for accurate metric recovery. By design, affine-invariant depth methods do not take the focal length information into account during training. As shown in Sec. 3.1, only from the image appearance, various focal lengths may cause metric ambiguity, thus they decouple the depth scale in training. To solve the problem of varying focal lengths, CamConv [38] encodes the camera model in the network, which enforces the network to implicitly understand camera models from the image appearance and then bridges the imaging size to the real-world size. However, training data contains limited images and types of cameras, which challenges data diversity and network capacity. We propose a canonical camera transformation method in training, inspired by the canonical pose space from human body reconstruction methods [39]. We transform all training data to a canonical camera space where the processed images are coarsely regarded as captured by the same camera. To achieve such transformation, we propose two different methods. The first one tries to adjust the image appearance to simulate the canonical camera, while the other one transforms the ground-truth labels for supervision. Camera models are not encoded in the network, making our method easily applicable to existing architectures. During inference, a de-canonical transformation is employed to recover metric information. To further boost the depth accuracy, we propose a random proposal normalization loss. It is inspired by the scale-shift invariant loss [25], [27], [29] decoupling the depth scale to emphasize the single image's distribution. However, they perform on the whole image, which inevitably squeezes the fine-grained depth difference. We propose to randomly crop several patches from images and enforce the scale-shift invariant loss [25], [27] on them. Our loss emphasizes the local geometry and distribution of the single image.

我们针对零样本度量深度和表面法线估计的挑战提出了有针对性的解决方案。对于度量尺度恢复，我们首先分析单目深度估计中的度量模糊问题，并研究深度中的不同相机参数，包括像素尺寸、焦距和传感器尺寸。我们观察到，焦距是准确进行度量恢复的关键因素。从设计上看，仿射不变深度方法在训练过程中不考虑焦距信息。如第 3.1 节所示，仅从图像外观来看，不同的焦距可能会导致度量模糊，因此它们在训练中解耦了深度尺度。为了解决焦距变化的问题，CamConv [38] 在网络中对相机模型进行编码，这迫使网络从图像外观中隐式理解相机模型，然后将成像尺寸与真实世界尺寸联系起来。然而，训练数据包含的图像和相机类型有限，这对数据多样性和网络容量提出了挑战。受人体重建方法中的规范姿态空间 [39] 的启发，我们在训练中提出了一种规范相机变换方法。我们将所有训练数据转换到一个规范相机空间，在这个空间中，处理后的图像可以粗略地视为由同一台相机拍摄的。为了实现这种变换，我们提出了两种不同的方法。第一种方法试图调整图像外观以模拟规范相机，而另一种方法则变换用于监督的真实标签。相机模型未在网络中编码，这使得我们的方法易于应用于现有架构。在推理过程中，采用反规范变换来恢复度量信息。为了进一步提高深度估计的准确性，我们提出了一种随机提议归一化损失。它受到尺度平移不变损失 [25]、[27]、[29] 的启发，该损失解耦了深度尺度以强调单张图像的分布。然而，它们是在整个图像上执行的，这不可避免地压缩了细粒度的深度差异。我们提议从图像中随机裁剪几个图像块，并对它们施加尺度平移不变损失 [25]、[27]。我们的损失强调单张图像的局部几何结构和分布。

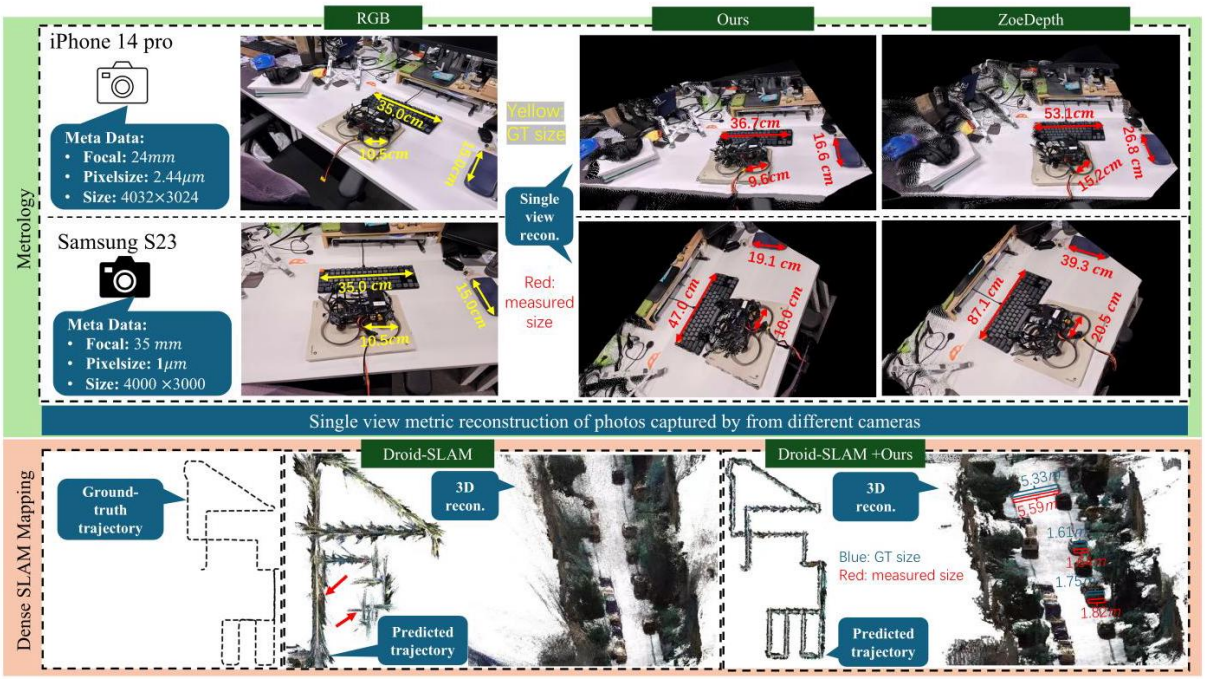

Fig. 3 - Top (metrology for a complex scene): we use two phones (iPhone 14 pro and Samsung Galaxy S23) to capture the scene and measure the size of several objects, including a drone which has never occurred in the whole training set. With the photos' metadata, we perform 3D metric reconstruction and then measure object sizes (marked in red), which are close to the ground truth (marked in yellow). Compared with ZoeDepth [36], our measured sizes are closer to ground truth. Bottom (dense SLAM mapping): existing SoTA mono-SLAM methods usually face scale drift problems (see the red arrows) in large-scale scenes and are unable to achieve the metric scale, while, naively inputting our metric depth, Droid-SLAM [37] can recover much more accurate trajectory and perform the metric dense mapping (see the red measurements). Note that all testing data are unseen to our model.

图 3 - 顶部 (复杂场景的度量): 我们使用两部手机 (iPhone 14 pro 和三星 Galaxy S23) 拍摄场景，并测量了几个物体的尺寸，其中包括一架在整个训练集中从未出现过的无人机。利用照片的元数据，我们进行了三维度量重建，然后测量了物体的尺寸 (用红色标记)，这些尺寸接近真实值 (用黄色标记)。与 ZoeDepth [36] 相比，我们测量的尺寸更接近真实值。底部 (密集 SLAM 建图): 现有的最先进的单目 SLAM 方法在大规模场景中通常会面临尺度漂移问题 (见红色箭头)，并且无法实现度量尺度，而简单地输入我们的度量深度，Droid - SLAM [37] 可以恢复更准确的轨迹并进行度量密集建图 (见红色测量值)。请注意，我们的模型从未见过所有测试数据。

For surface normal, the biggest challenge is the lack of diverse (outdoor) annotations. Compared to reconstruction-based annotation methods [35], [40], directly producing normal labels from network-predicted depth is more efficient and scalable. The quality of such pseudo-normal labels, however, is bounded by the accuracy of the depth network. Fortunately, we observe that robust metric depth models are scalable geometric learners, containing abundant information for normal estimation. Weak supervision from the pseudo normal annotations transformed by learned metric depth can effectively prevent the normal estimator from collapsing caused by GT absence. Furthermore, this supervision can guide the normal estimator to generalize on large-scale unlabeled data. Based on such observation, we propose a joint depth-normal optimization module to distill knowledge from diverse depth datasets. During optimization, our normal estimator learns from three sources: (1) Groundtruth normal labels, though they are much fewer compared to depth annotations (2) An explicit learning objective to constrain depth-normal consistency. (3) Implicit and thorough knowledge transfer from depth to normal through feature fusion, which is more tolerant to unsatisfactory initial prediction than the explicit counterparts [41], [42]. To achieve this, we implement the optimization module using deep recurrent blocks. While previous researchers have employed similar recurrent modules to optimize depth [42], [43], [44], disparity [45], ego-motion [37], or optical flows [46], it is the first time that normal is iteratively optimized together with depth in a learning-based scheme. Benefiting from the joint optimization module, our models can efficiently learn normal knowledge from large-scale depth datasets even without labels.

对于表面法线估计，最大的挑战是缺乏多样化的 (室外) 标注。与基于重建的标注方法 [35]、[40] 相比，直接从网络预测的深度生成法线标签更高效且可扩展。然而，这种伪法线标签的质量受到深度网络准确性的限制。幸运的是，我们观察到鲁棒的度量深度模型是可扩展的几何学习者，包含了用于法线估计的丰富信息。由学习到的度量深度转换而来的伪法线标注的弱监督可以有效防止法线估计器因缺乏真实标签而崩溃。此外，这种监督可以引导法线估计器在大规模未标注数据上进行泛化。基于这一观察，我们提出了一个联合深度 - 法线优化模块，以从多样化的深度数据集中提炼知识。在优化过程中，我们的法线估计器从三个来源学习:(1) 真实法线标签，尽管与深度标注相比数量要少得多；(2) 一个明确的学习目标，用于约束深度 - 法线的一致性；(3) 通过特征融合从深度到法线的隐式和全面的知识转移，与显式方法 [41]、[42] 相比，它对不理想的初始预测更具容忍性。为了实现这一点，我们使用深度循环块实现了优化模块。虽然之前的研究人员已经使用类似的循环模块来优化深度 [42]、[43]、[44]、视差 [45]、自我运动 [37] 或光流 [46]，但这是首次在基于学习的方案中同时迭代优化法线和深度。受益于联合优化模块，我们的模型即使在没有标签的情况下也能从大规模深度数据集中高效地学习法线知识。

With the proposed method, we can stably scale up model training to 16 million images from 16 datasets of diverse scene types (indoor and outdoor, real or synthetic data), camera models (tens of thousands of different cameras), and annotation categories (with or without normal), leading to zero-shot transferability and significantly improved accuracy. Fig. 4 illustrates how the large-scale data with depth annotations directly facilitate metric depth and surface normal learning. The metric depth and normal given by our model directly broaden the applications

in downstream tasks. We achieve state-of-the-art performance on over 16 depth and normal benchmarks, see Fig. 1. Our model can accurately reconstruct metric 3D from randomly collected Internet images, enabling plausible single-image metrology. For examples (Fig. 3), we recovery real-world metric to improve monocular SLAM [37], [47] and facilitate large-scale 3D reconstruction [48]. Our main contributions can be summarized as:

通过所提出的方法，我们可以将模型训练稳定地扩展到来自 16 个不同场景类型 (室内和室外、真实或合成数据)、相机型号 (数以万计的不同相机) 和标注类别 (有或没有法线) 的 1600 万张图像，从而实现零样本可迁移性并显著提高准确率。图 4 展示了带有深度标注的大规模数据如何直接促进度量深度和表面法线学习。我们的模型给出的度量深度和法线直接拓宽了下游任务的应用范围。我们在 16 个以上的深度和法线基准测试中达到了最先进的性能，见图 1。我们的模型可以从随机收集的互联网图像中准确重建度量 3D，实现合理的单图像计量。例如 (图 3)，我们恢复现实世界的度量以改进单目 SLAM [37]、[47] 并促进大规模 3D 重建 [48]。我们的主要贡献可以总结如下:

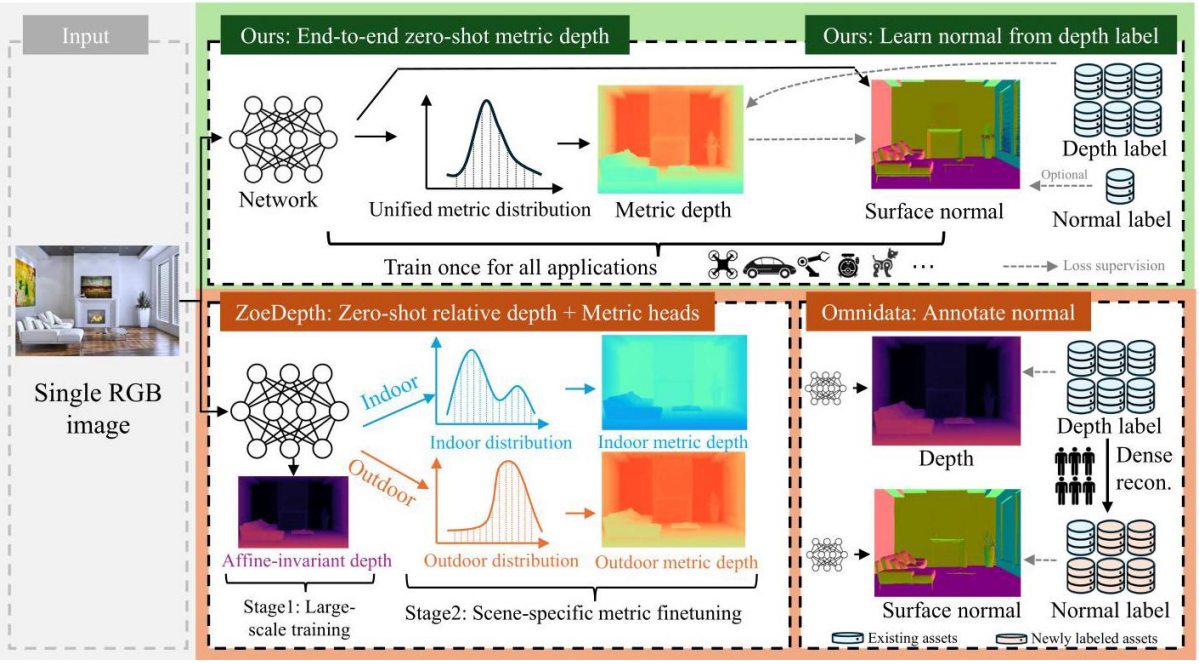

Fig. 4 –Overall methodology. Our method takes a single image to predict the metric depth and surface normal simultaneously. We apply large-scale data training directly for metric depth estimation rather than affine invariant depth, enabling end-to-end zero-shot metric depth estimation for various applications using a single model. For normals, we enable learning from depth labels only, alleviating the demand for dense reconstruction to generate large-scale normal labels.

图 4 –整体方法。我们的方法采用单张图像同时预测度量深度和表面法线。我们直接应用大规模数据训练进行度量深度估计，而不是仿射不变深度估计，从而使用单个模型为各种应用实现端到端的零样本度量深度估计。对于法线，我们实现了仅从深度标签中学习，减轻了通过密集重建生成大规模法线标签的需求。

- We propose a canonical camera transformation method to address metric depth ambiguity across different camera settings. This approach facilitates training zero-shot monocular metric depth models using large-scale datasets.

- 我们提出了一种规范相机变换方法，以解决不同相机设置下的度量深度模糊性问题。这种方法有助于使用大规模数据集训练零样本单目度量深度模型。

- We design a random proposal normalization loss to effectively improve metric depth.

- 我们设计了一种随机提议归一化损失，以有效提高度量深度。

- We propose a joint depth-normal optimization module to learn normal on large-scale datasets without normal annotation, distilling knowledge from the metric depth estimator.

- 我们提出了一个联合深度 - 法线优化模块，用于在没有法线标注的大规模数据集上学习法线，从度量深度估计器中提炼知识。

- Our models rank 1st on a wide variety of depth and surface normal benchmarks. It can perform high-quality 3D metric structure recovery in the wild and benefit several downstream tasks, such as mono-SLAM [37], [49], 3D scene reconstruction [48], and metrology [50].

- 我们的模型在各种深度和表面法线基准测试中排名第一。它可以在自然场景中进行高质量的 3D 度量结构恢复，并有益于多个下游任务，如单目 SLAM [37]、[49]、3D 场景重建 [48] 和计量学 [50]。

2 RELATED WORK

2 相关工作

3D reconstruction from a single image. The reconstruction of diverse objects from a singular image has been extensively investigated in prior research [51], [52], [53]. These methodologies exhibit proficiency in generating high-fidelity 3D models encompassing various entities such as cars, planes, tables, and human bodies [54], [55]. The primary challenge lies in optimizing the recovery of object details, devising efficient representations within constrained memory resources, and achieving generalization across a broader spectrum of objects. However, these approaches typically hinge upon learning object-specific or instance-specific priors, often derived from 3D supervision, thereby rendering them unsuitable for comprehensive scene reconstruction. In addition to the aforementioned efforts on object reconstruction, several studies focus on scene reconstruction from single images. Saxena et al. [56] adopt an approach that segments the entire scene into multiple small planes, with the 3D structure represented based on the orientation and positioning of these planes. More recently, LeReS [25] proposed employing a robust monocular depth estimation model for scene reconstruction. Nonetheless, their method is limited to recovering shapes up to a certain scale. Zhang et al. [57] recently introduced a zero-shot geometry-preserving depth estimation model capable of providing depth predictions up to an unknown scale. In contrast to the aforementioned methodologies, our approach excels in recovering the metric 3D structure of scenes.

单图像 3D 重建。从单张图像重建各种物体的问题在先前的研究中得到了广泛探讨 [51]、[52]、[53]。这些方法能够熟练生成涵盖各种实体 (如汽车、飞机、桌子和人体) 的高保真 3D 模型 [54]、[55]。主要挑战在于优化物体细节的恢复、在有限的内存资源内设计有效的表示方式以及在更广泛的物体范围内实现泛化。然而，这些方法通常依赖于学习特定物体或特定实例的先验知识，这些知识往往来自 3D 监督，因此不适合进行全面的场景重建。除了上述关于物体重建的工作外，一些研究专注于从单张图像进行场景重建。Saxena 等人 [56] 采用了一种将整个场景分割成多个小平面的方法，3D 结构基于这些平面的方向和位置来表示。最近，LeReS [25] 提出使用一种强大的单目深度估计模型进行场景重建。然而，他们的方法仅限于在一定尺度上恢复形状。Zhang 等人 [57] 最近引入了一种零样本保几何深度估计模型，能够提供未知尺度的深度预测。与上述方法相比，我们的方法在恢复场景的度量 3D 结构方面表现出色。

Supervised monocular depth estimation. Following the establishment of several benchmarks [58], [59], neural network-based methods [17], [19], [30] have dominated this task. These approaches often regress continuous depth by aggregating the information from an image [60]. However, since depth distribution varies significantly with different RGB values, some methods tend to discretize depth and reformulate the problem as a classification task [18] for better performance. The generalization of deep models for 3D metric recovery faces two challenges: adapting to diverse scenes and predicting accurate metric information under various camera settings. Recent methods [18], [21], [22] have effectively addressed the first challenge by creating large-scale relative depth datasets like DIW [24] and OASIS [23] to learn relative relations, which lose geometric structure information. To enhance geometry, methods like MiDaS [27], LeReS [25], and HDN [29] employ affine-invariant depth learning. These approaches, utilizing large-scale data, have continuously improved performance and scene generalization. However, they inherently struggle to recover metric information. Thus, achieving both strong generalization and accurate metric data across diverse scenes remains a key challenge to be addressed. Con-currently, ZoeDepth [36], ZeroDepth [61], and UniDepth [62] apply varying strategies to tackle this challenge.

有监督的单目深度估计。在建立了几个基准测试 [58]、[59] 之后，基于神经网络的方法 [17]、[19]、[30] 在这项任务中占据了主导地位。这些方法通常通过聚合图像信息来回归连续深度 [60]。然而，由于深度分布随不同的 RGB 值有显著变化，一些方法倾向于将深度离散化，并将问题重新表述为分类任务 [18] 以获得更好的性能。用于 3D 度量恢复的深度模型的泛化面临两个挑战: 适应不同的场景和在各种相机设置下预测准确的度量信息。最近的方法 [18]、[21]、[22] 通过创建像 DIW [24] 和 OASIS [23] 这样的大规模相对深度数据集来学习相对关系，有效解决了第一个挑战，但这些方法丢失了几何结构信息。为了增强几何信息，像 MiDaS [27]、LeReS [25] 和 HDN [29] 这样的方法采用仿射不变深度学习。这些利用大规模数据的方法不断提高了性能和场景泛化能力。然而，它们本质上难以恢复度量信息。因此，在不同场景下同时实现强泛化能力和准确的度量数据仍然是一个需要解决的关键挑战。同时，ZoeDepth [36]、ZeroDepth [61] 和 UniDepth [62] 采用了不同的策略来应对这一挑战。

Surface normal estimation. Compared to metric depth, surface normal suffers no metric ambiguity and preserves local geometry better. These properties attract researchers to apply normal in various vision tasks like localization [11], mapping [63], and 3D scene reconstruction [6], [64]. Currently, learning-based methods [32], [33], [34], [42], [64], [65], [66], [67], [68], [69], [70], [71], [72] have dominated monocular surface normal estimation. Since normal labels required for training cannot be directly captured by sensors, previous works use [41], [58], [65], [67] kernel functions to annotate normal from dense indoor depth maps [58]. These annotations become incomplete on reflective surfaces and inaccurate at object boundaries. To learn from such imperfect annotations, GeoNet [41] proposes to enforce depth-normal consistency with mutual transformation modules, ASN [71], [72]

propose a novel adaptive surface normal constraint to facilitate joint depth-normal learning, and Bae et al. [33] propose an uncertainty-based learning objective. Nonetheless, it is challenging for such methods to further increase their generalization, due to the limited dataset size and the diversity of scenes, especially for outdoor scenarios. Omni-data [35] advances to fill this gap by building 1300M frames of normal annotation. Normal-in-the-wild [73] proposes a pipeline for efficient normal labeling. A con-current work DSINE [74] also employs diverse datasets to train generalizable surface normal estimators. However, further scaling up normal labels remains difficult. This underscores research significance in finding an efficient way to distill prior from other types of annotation.

表面法线估计。与度量深度相比，表面法线不存在度量模糊性，并且能更好地保留局部几何信息。这些特性吸引研究人员将法线应用于各种视觉任务，如定位 [11]、建图 [63] 和三维场景重建 [6]、[64]。目前，基于学习的方法 [32]、[33]、[34]、[42]、[64]、[65]、[66]、[67]、[68]、[69]、[70]、[71]、[72] 在单目表面法线估计中占据主导地位。由于训练所需的法线标签无法由传感器直接捕获，以往的工作 [41]、[58]、[65]、[67] 使用核函数从密集的室内深度图 [58] 中标注法线。这些标注在反射表面上会变得不完整，在物体边界处也不准确。为了从这种不完美的标注中学习，GeoNet[41] 提出使用相互转换模块来强制深度 - 法线一致性，ASN[71]、[72] 提出了一种新颖的自适应表面法线约束以促进深度 - 法线联合学习，Bae 等人 [33] 提出了一种基于不确定性的学习目标。尽管如此，由于数据集规模有限和场景的多样性，尤其是在室外场景中，此类方法要进一步提高其泛化能力仍具有挑战性。Omni - data[35] 通过构建 1300M 帧的法线标注来填补这一空白。Normal - in - the - wild[73] 提出了一种高效的法线标注流程。同期工作 DSINE[74] 也采用多样化的数据集来训练具有泛化能力的表面法线估计器。然而，进一步扩大法线标签的规模仍然困难。这凸显了寻找一种从其他类型标注中提取先验信息的有效方法的研究意义。

Deep iterative refinement for geometry. Iterative refinement enables multi-step coarse-to-fine prediction and benefits a wide range of geometry estimation tasks, such as optical flow estimation [46], [75], [76], depth completion [43], [77], [78], and stereo matching [?], [45], [79]. Classical iterative refinements [75], [77] optimize directly on high-resolution outputs using high-computing-cost operators, limiting researchers from applying more iterations for better predictions. To address this limitation, RAFT [46] proposes to optimize an intermediate low-resolution prediction using ConvGRU modules. For monocular depth estimation, IEBins [44] employs similar methods to optimize depth-bin distribution. Differently, IronDepth [42] propagates depth on pre-computed local surfaces. Regarding surface normal refinement, Lenssen et al. [80] propose a deep iterative method to optimize normal from point clouds. Zhao et al. [81] design a solver to refine depth and normal jointly, but it requires multi-view prior and per-sample post optimization. Without multi-view prior, such a non-learnable optimization method could fail due to unsatisfactory initial predictions. All the monocular methods [42], [44], [80], however, iterate over either depth or normal independently. In contrast, our joint optimization module tightly couples depth and normal with each other.

几何的深度迭代细化。迭代细化能够实现多步从粗到精的预测，并且有益于广泛的几何估计任务，如光流估计 [46]、[75]、[76]、深度补全 [43]、[77]、[78] 和立体匹配 [?]、[45]、[79]。经典的迭代细化方法 [75]、[77] 使用高计算成本的算子直接在高分辨率输出上进行优化，限制了研究人员为获得更好的预测而进行更多次迭代。为了解决这一限制，RAFT[46] 提出使用卷积门控循环单元 (ConvGRU) 模块对中间低分辨率预测进行优化。对于单目深度估计，IEBins[44] 采用类似的方法来优化深度区间分布。不同的是，IronDepth[42] 在预计算的局部表面上传播深度。关于表面法线细化，Lenssen 等人 [80] 提出了一种深度迭代方法，用于从点云中优化法线。Zhao 等人 [81] 设计了一个求解器来联合细化深度和法线，但它需要多视图先验信息和每个样本的后优化。在没有多视图先验信息的情况下，这种不可学习的优化方法可能会因初始预测不理想而失败。然而，所有单目方法 [42]、[44]、[80] 都只是独立地对深度或法线进行迭代。相比之下，我们的联合优化模块将深度和法线紧密地耦合在一起。

Large-scale data training. Recently, various natural language problems and computer vision problems [82], [83], [84] have achieved impressive progress with large-scale data training. CLIP [83] is a promising classification model trained on billions of paired image-language data pairs, achieving achieves state-of-the-art performance zero-shot classification benchmarks. Dinov2 [85] collects 142M images to conduct vision-only self-supervised learning for vision transformers [86]. Generative models like LDM [87] have also undergone billion-level data pre-training. For depth prediction, large-scale data training has been widely applied. Ranft et al. [27] mix over 2 million data in training, LeReS [26] collects over 300 thousands data, Eftekhar et al. [35] also merge millions of data to build a strong depth prediction model. To train a zero-shot surface normal estimator, Omni-data [35] performs dense reconstruction to generate 14M frames with surface normal annotations.

大规模数据训练。最近，各种自然语言问题和计算机视觉问题 [82]、[83]、[84] 通过大规模数据训练取得了显著进展。CLIP[83] 是一个很有前景的分类模型，它在数十亿对图像 - 语言数据对上进行训练，在零样本分类基准测试中取得了最先进的性能。Dinov2[85] 收集了 142M 张图像，对视觉变换器 [86] 进行仅视觉的自监督学习。像潜在扩散模型 (LDM)[87] 这样的生成模型也进行了数十亿级别的数据预训练。对于深度预测，大规模数据训练已被广泛应用。Ranft 等人 [27] 在训练中混合了超过 200 万个数据，LeReS[26] 收集了超过 30 万个数据，Eftekhar 等人 [35] 也合并了数百万个数据来构建一个强大的深度预测模型。为了训练一个零样本表面法线估计器，Omni - data[35] 进行密集重建，以生成带有表面法线标注的 14M 帧。

3 Method

3 方法

Preliminaries. We consider the pin-hole camera model with intrinsic parameters formulated as: $\left[\left[\widehat{f}/\delta, 0, u_0\right], \left[0, \widehat{f}/\delta, v_0\right]\right]$, $[0, 0, 1]]$, where $\widehat{f}$ is the focal length (in micrometers), δ is the pixel size (in micrometers), and (u_0, v_0) is the principle center. $f = \widehat{f}/\delta$ is the pixel-represented focal length.

预备知识。我们考虑内参公式如下的针孔相机模型: $\left[\left[\widehat{f}/\delta, 0, u_0\right], \left[0, \widehat{f}/\delta, v_0\right]\right]$ ， $[0, 0, 1]]$ ， 其中 $\widehat{f}$ 是焦距 (单位: 微米)，δ 是像素尺寸 (单位: 微米)，(u_0, v_0) 是主点。$f = \widehat{f}/\delta$ 是以像素表示的焦距。

3.1 Ambiguity Issues in Metric Depth Estimation

3.1 度量深度估计中的歧义问题

Figure 5 illustrates an instance of photographs captured using diverse cameras and at varying distances. Solely based on visual inspection, one might erroneously infer that the last two images originate from a comparable location and are captured by the same camera. However, due to differing focal lengths, these images are indeed captured from distinct locations. Consequently, accurate knowledge of camera intrinsic parameters becomes imperative for metric estimation from a single image; otherwise, the problem becomes ill-posed. Recent methodologies, such as MiDaS [27] and LeReS [25], mitigate this metric ambiguity by decoupling metric estimation from direct supervision and instead prioritize learning affine-invariant depth.

图 5 展示了使用不同相机在不同距离拍摄的照片示例。仅通过肉眼观察，人们可能会错误地推断最后两张图像来自相近的位置，并且是由同一台相机拍摄的。然而，由于焦距不同，这些图像实际上是从不同位置拍摄的。因此，准确了解相机内参对于单张图像的度量估计至关重要；否则，该问题将是病态的。最近的方法，如 MiDaS [27] 和 LeReS [25]，通过将度量估计与直接监督解耦，转而优先学习仿射不变深度，来缓解这种度量歧义。

Figure 6 (A) depicts the pin-hole perspective projection, where object A located at distance d_a is projected to A'. Adhering to the principle of similarity, it is obvious that:

图 6 (A) 描绘了针孔透视投影，其中位于距离 d_a 处的物体 A 被投影到 A' 处。根据相似原理，显然有:

$$d_a = \widehat{S}\left[\frac{\widehat{f}}{\widehat{S}'}\right] = \widehat{S}\cdot\alpha \tag{1}$$

where $\widehat{S}$ and $\widehat{S}'$ are the real and imaging size respectively. The symbol " signifies that variables are expressed in physical metrics (e.g., millimeters). To ascertain d_a from a single image, one must have access to the focal length, imaging size of the object, and real-world object size. Estimating the focal length from a single image is challenging and inherently ill-posed. Despite numerous methods having been explored [25], [88], achieving satisfactory accuracy remains elusive. Hereby, we simplify the scenario by assuming known focal lengths for the training/test images. In contrast, understanding the imaging size is much easier for a neural network. To obtain the real-world object size, a neural network needs to understand the semantic scene layout and the object, at which a neural network excels. We define $\alpha = \widehat{f}/\widehat{S}'$, showing that d_a is proportional to α.

其中 $\widehat{S}$ 和 $\widehat{S}'$ 分别是真实尺寸和成像尺寸。符号 "" 表示变量以物理度量 (例如，毫米) 表示。要从单张图像中确定 d_a，必须知道焦距、物体的成像尺寸和现实世界中物体的尺寸。从单张图像估计焦距具有挑战性，并且本质上是病态的。尽管已经探索了许多方法 [25]、[88]，但仍难以达到令人满意的精度。因此，我们通过假设训练/测试图像的焦距已知来简化场景。相比之下，神经网络更容易理解成像尺寸。为了获得现实世界中物体的尺寸，神经网络需要理解语义场景布局和物体，而这正是神经网络擅长的。我们定义 $\alpha = \widehat{f}/\widehat{S}'$，表明 d_a 与 α 成正比。

Fig. 5 - Photos of a chair captured at different distances with different cameras. The first two photos are captured at the same distance but with different cameras, while the last one is taken at a closer distance with the same camera as the first one.

图 5 - 使用不同相机在不同距离拍摄的椅子照片。前两张照片是在相同距离但使用不同相机拍摄的，而最后一张是使用与第一张相同的相机在更近的距离拍摄的。

We observe the following regarding sensor size, pixel size, and focal length.

我们对传感器尺寸、像素尺寸和焦距有以下观察。

O1: Sensor size and pixel size do not affect metric depth estimation. Based on perspective projection (Fig. 6 (A)), sensor size only influences the field of view (FOV) and is not relevant to α , hence it does not affect metric depth estimation. For pixel size, consider two cameras with different pixel sizes $(\delta_1 = 2\delta_2)$ but the same focal length $\widehat{f}$, capturing the same object at distance d_a . Fig. 6 (B) displays their captured images. According to the preliminaries, the pixel-represented focal length is $f_1 = \frac{1}{2} f_2$. Since the second camera has smaller pixel sizes, the resolution of the pixel-represented image is given by $S_1' = \frac{1}{2} S_2'$, despite both having the same projected imaging size $\widehat{S}'$. According to Eq. (1), we have $\frac{\widehat{f}}{\delta_1 \cdot S_1'} = \frac{\widehat{f}}{\delta_2 \cdot S_2'}$, which implies $\alpha_1 = \alpha_2$ and $d_1 = d_2$. This means that variations in camera sensors do not impact the estimation of metric depth.

O1: 传感器尺寸和像素尺寸不影响度量深度估计。基于透视投影 (图 6 (A))，传感器尺寸仅影响视场 (FOV)，与 α 无关，因此它不影响度量深度估计。对于像素尺寸，考虑两个像素尺寸不同 ($(\delta_1 = 2\delta_2)$) 但焦距相同 ($\widehat{f}$) 的相机，在距离 d_a 处拍摄同一物体。图 6 (B) 展示了它们拍摄的图像。根据预备知识，以像素表示的焦距为 $f_1 = \frac{1}{2} f_2$ 。由于第二个相机的像素尺寸较小，尽管两者的投影成像尺寸相同 ($\widehat{S}'$)，但以像素表示的图像分辨率由 $S_1' = \frac{1}{2} S_2'$ 给出。根据公式 (1)，我们有 $\frac{\widehat{f}}{\delta_1 \cdot S_1'} = \frac{\widehat{f}}{\delta_2 \cdot S_2'}$ ，这意味着 $\alpha_1 = \alpha_2$ 和 $d_1 = d_2$ 。这意味着相机传感器的变化不会影响度量深度的估计。

O2: The focal length is vital for metric depth estimation. Figure 5 shows the challenge of metric ambiguity caused by an unspecified focal length, which is further discussed in Figure 7. In the scenario where two cameras $\left(\widehat{f}_1 = 2\widehat{f}_2\right)$ are positioned at distances $d_1 = 2d_2$, the imaging sizes remain consistent for both cameras. As a result, the neural network struggles to distinguish between different supervision labels based solely on visual cues. To address this issue, we propose a canonical camera transformation method to reduce conflicts between supervision requirements and image representations.

O2: 焦距对于度量深度估计至关重要。图 5 展示了由未指定焦距引起的度量模糊问题，图 7 进一步讨论了该问题。在两个相机 ($\left(\widehat{f}_1 = 2\widehat{f}_2\right)$) 分别位于距离 ($d_1 = 2d_2$) 的场景中，两个相机的成像尺寸保持一致。因此，神经网络仅根据视觉线索难以区分不同的监督标签。为了解决这个问题，我们提出了一种标准相机变换方法，以减少监督要求和图像表示之间的冲突。

Fig. 6 - Pinhole camera model. (A)Object A positioned at a distance d_a undergoes projection onto the image plane. (B) Employing two cameras for capturing an image of the car. The left one has a larger pixel size. Although the projected imaging sizes are the same, the pixel-represented images (resolution) are different.

图 6 - 针孔相机模型。(A) 位于距离 d_a 处的物体 (A) 投影到图像平面上。(B) 使用两个相机拍摄汽车的图像。左边的相机像素尺寸较大。虽然投影成像尺寸相同，但以像素表示的图像 (分辨率) 不同。

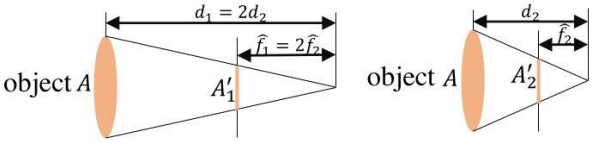

Fig. 7 - Illustration of two cameras with different focal length at different distance. As $f_1 = 2f_2$ and $d_1 = 2d_2$, A is projected to two image planes with the same imaging size (i.e. $A'_1 = A'_2$).

图 7 - 不同焦距的两个相机在不同距离的示意图。由于 $f_1 = 2f_2$ 和 $d_1 = 2d_2$, A 投影到两个成像尺寸相同的图像平面上 (即 $A'_1 = A'_2$)。

Unlike depth, surface normal does not have any metric ambiguity problem. In Fig. 8, we illustrate this concept with two depth maps at varying scales, denoted as $\mathbf{D}_1$ and $\mathbf{D}_2$, featuring distinct metrics d1 and d2 , respectively, where $d_1 < d_2$. After upprojecting the depth to the 3D point cloud, the dolls are in different depths d_1 and d_2 . However, the surface normals $\mathbf{n}_1$ and $\mathbf{n}_2$ corresponding to a certain pixel $A' \in \mathbf{I}$ remain the same.

与深度不同，表面法线没有任何度量模糊问题。在图 8 中，我们用两个不同尺度的深度图 (分别表示为 $\mathbf{D}_1$ 和 $\mathbf{D}_2$ ，具有不同的度量 d1 和 d2 ，其中 $d_1 < d_2$) 来说明这一概念。将深度反投影到三维点云后，玩偶处于不同的深度 (d_1 和 d_2)。然而，对应于某个像素 ($A' \in \mathbf{I}$) 的表面法线 ($\mathbf{n}_1$ 和 $\mathbf{n}_2$) 保持不变。

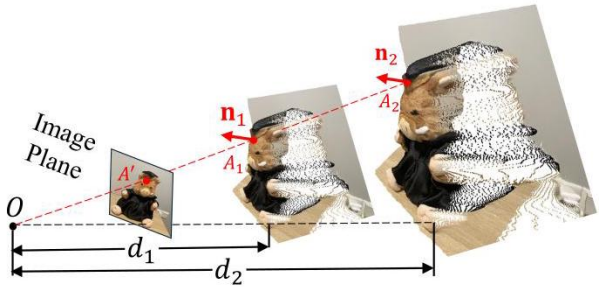

Fig. 8 - The metric-agnostic property of normal. With differently predicted metrics d_1 and d_2 , the pixel A' on the image will be back-projected to 3D points A_1 and A_2 , respectively. The surface normal $\mathbf{n}_1$ at A_1 and $\mathbf{n}_2$ at A_2 remain the same.

图 8 - 法线的度量无关特性。使用不同预测的度量 (d_1 和 d_2)，图像上的像素 (A') 将分别反投影到 3D 个点 (A_1 和 A_2)。在 A_1 处的表面法线 ($\mathbf{n}_1$) 和在 A_2 处的表面法线 ($\mathbf{n}_2$) 保持不变。

3.2 Canonical Camera Transformation

3.2 标准相机变换

The fundamental concept entails establishing a canonical camera space $\left(\left(f_x^c, f_y^c\right)\right.$, with $f_x^c = f_y^c = f^c$ in experimental settings) and transposing all training data into this designated space. Consequently, all data can be broadly construed as being captured by the canonical camera. We propose two transformation methods, i.e. either transforming the input image $\left(\mathbf{I} \in \mathbb{R}^{H \times W \times 3}\right)$ or the ground-truth (GT) label $\left(\mathbf{D} \in \mathbb{R}^{H \times W}\right)$. The initial intrinsics are $\{f, u_0, v_0\}$.

基本概念包括在实验设置中建立一个规范相机空间 $\left(\left(f_x^c, f_y^c\right)\right.$, with $f_x^c = f_y^c = f^c$, 并将所有训练数据转换到这个指定空间。因此，所有数据都可以广义地理解为是由规范相机捕获的。我们提出了两种转换方法，即要么转换输入图像 $\left(\mathbf{I} \in \mathbb{R}^{H \times W \times 3}\right)$ ，要么转换真实标签 (GT) $\left(\mathbf{D} \in \mathbb{R}^{H \times W}\right)$ 。初始内参为 $\{f, u_0, v_0\}$ 。

Method1: transforming depth labels (CSTM_label). Fig. 5's ambiguity is for depths. Consequently, our initial approach directly addresses this issue by transforming the ground-truth depth labels. Specifically, we rescale the ground-truth depth $(\mathbf{D}^*)$ using the ratio $\omega_d = \frac{f^c}{f}$ during training, denoted as $\mathbf{D}_c = \omega_d \mathbf{D}^*$. The original camera model undergoes transformation to f^c, u_0, v_0 . In inference, the predicted depth $(\mathbf{D}_c)$ exists in the canonical space and necessitates a de-canonical transformation to restore metric information, expressed as $\mathbf{D} = \frac{1}{\omega_d}\mathbf{D}_c$. It is noteworthy that the input $\mathbf{I}$ remains unaltered, represented as $\mathbf{I}_c = \mathbf{I}$.

方法 1: 转换深度标签 (CSTM_label)。图 5 的模糊性与深度有关。因此，我们的第一种方法通过转换真实深度标签直接解决了这个问题。具体来说，我们在训练期间使用比率 $\omega_d = \frac{f^c}{f}$ 对真实深度 $(\mathbf{D}^*)$ 进行重新缩放，记为 $\mathbf{D}_c = \omega_d \mathbf{D}^*$ 。原始相机模型转换为 f^c, u_0, v_0 。在推理过程中，预测深度 $(\mathbf{D}_c)$ 存在于规范空间中，需要进行反规范转换以恢复度量信息，表示为 $\mathbf{D} = \frac{1}{\omega_d}\mathbf{D}_c$ 。值得注意的是，输入 $\mathbf{I}$ 保持不变，表示为 $\mathbf{I}_c = \mathbf{I}$ 。

Method2: transforming input images (CSTM_image). From an alternate perspective, the ambiguity arises due to the resemblance in image appearance. Consequently, this methodology aims to alter the input image to emulate the imaging effects of the canonical camera. Specifically, the image $\mathbf{I}$ undergoes resizing using the ratio $\omega_r = \frac{f^c}{f}$, denoted as $\mathbf{I}c = \mathcal{T}(\mathbf{I}, \omega_r)$, where $\mathcal{T}(\cdot)$ signifies image resizing. As a result of resizing the optical center, the canonical camera model becomes $f^c, \omega_r u_0, \omega_r v_0$. The ground-truth labels are resized without scaling, represented as $\mathbf{D}_c^* = \mathcal{T}(\mathbf{D}^*, \omega_r)$. In inference, the de-canonical transformation involves resizing the prediction to its original dimensions without scaling, expressed as $\mathbf{D} = \mathcal{T}\left(\mathbf{D}_c, \frac{1}{\omega_r}\right)$.

方法 2: 转换输入图像 (CSTM_image)。从另一个角度来看，模糊性是由于图像外观的相似性引起的。因此，这种方法旨在改变输入图像以模拟规范相机的成像效果。具体来说，图像 $\mathbf{I}$ 使用比率 $\omega_r = \frac{f^c}{f}$ 进行调整大小，记为 $\mathbf{I}c = \mathcal{T}(\mathbf{I}, \omega_r)$，其中 $\mathcal{T}(\cdot)$ 表示图像调整大小。由于调整了光学中心的大小，规范相机模型变为 $f^c, \omega_r u_0, \omega_r v_0$。真实标签在不进行缩放的情况下进行调整大小，表示为 $\mathbf{D}_c^* = \mathcal{T}(\mathbf{D}^*, \omega_r)$。在推理过程中，反规范转换包括在不进行缩放的情况下将预测结果调整回其原始尺寸，表示为 $\mathbf{D} = \mathcal{T}\left(\mathbf{D}_c, \frac{1}{\omega_r}\right)$。

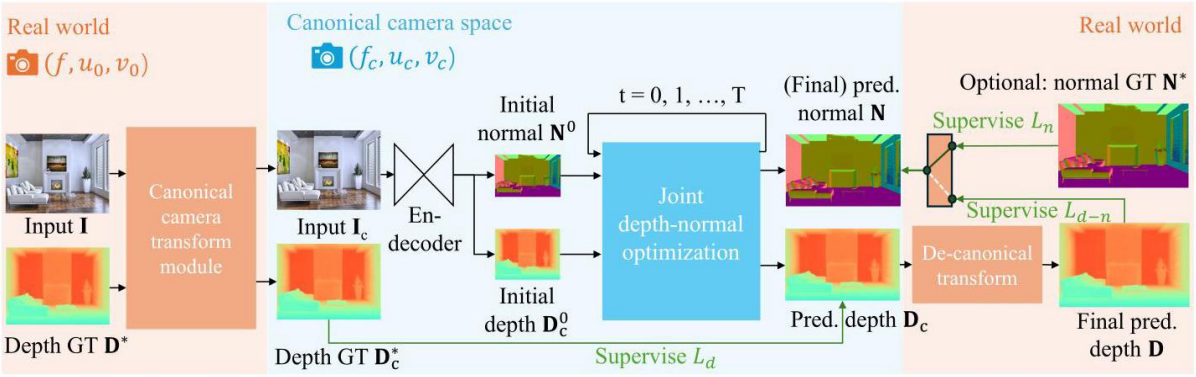

Fig. 9 - Pipeline. Given an input image I, we first transform it to the canonical space using CSTM. The transformed image I_c is fed into a standard depth-normal estimation model to produce the predicted metric depth D_c in the canonical space and metric-agnostic surface normal N. During training, D_c is supervised by a GT depth D_c^* which is also transformed into the canonical space. In inference, after producing the metric depth D_c in the canonical space, we perform a de-canonical transformation to convert it back to the space of the original input I. The canonical space transformation and de-canonical transformation are executed using camera intrinsics. The predicted normal N is supervised by depth-normal consistency via the recovered metric depth D as well as GT normal N^*, if available.

图 9 - 流程。给定输入图像 I，我们首先使用 CSTM 将其转换到规范空间。转换后的图像 I_c 被输入到一个标准的深度 - 法线估计模型中，以在规范空间中生成预测的度量深度 D_c 和与度量无关的表面法线 N。在训练期间，D_c 由也被转换到规范空间的真实深度 D_c^* 进行监督。在推理过程中，在规范空间中生成度量深度 D_c 后，我们进行反规范转换以将其转换回原始输入 I 的空间。规范空间转换和反规范转换使用相机内参执行。如果有可用的真实法线 N^*，预测的法线 N 通过恢复的度量深度 D 以及真实法线进行深度 - 法线一致性监督。

While similar transformations have been employed in MPSD [89] to normalized depth prediction, our approaches apply these modules to predict metric depth directly.

虽然在多尺度金字塔立体匹配 (MPSD)[89] 中已采用类似变换进行归一化深度预测，但我们的方法将这些模块用于直接预测度量深度。

Figure 9 shows the pipeline. After adopting either transformation, a patch is randomly cropped for training purposes. This cropping operation solely adjusts the field of view (FOV) and the optical center, thus averting any potential metric ambiguity issues. In the labels transformation approach, $\omega_r = 1$ and $\omega_d = \frac{f^c}{f}$, while in the images transformation method, $\omega_d = 1$ and $\omega_r = \frac{f^c}{f}$. Throughout the training process, the transformed ground-truth depth

labels $\mathbf{D}_c^*$ are employed as supervision. Importantly, since surface normals are not susceptible to metric ambiguity, no transformation is applied to normal labels $\mathbf{N}^*$.

图 9 展示了该流程。采用任一种变换后，会随机裁剪一个图像块用于训练。这种裁剪操作仅调整视场角 (FOV) 和光学中心, 从而避免任何潜在的度量模糊问题。在标签变换方法中, $\omega_r = 1$ 和 $\omega_d = \frac{f^c}{f}$ ，而在图像变换方法中，$\omega_d = 1$ 和 $\omega_r = \frac{f^c}{f}$ 。在整个训练过程中，使用变换后的真实深度标签 $\mathbf{D}_c^*$ 作为监督。重要的是，由于表面法线不受度量模糊的影响，因此不对法线标签 $\mathbf{N}^*$ 进行任何变换。

3.3 Jointly optimizing depth and normal

3.3 联合优化深度和法线

We propose to optimize metric depth and surface normal jointly in an end-to-end manner. This optimization is primarily aimed at leveraging a large amount of annotation knowledge available in depth datasets to improve normal estimation, particularly in outdoor scenarios where depth datasets contain significantly more annotations than normal datasets. In our experiments, we collect from the community 9488 K images with depth annotations across 14 outdoor datasets while less than 20 K outdoor normal-labeled images, presented in Tab. 5.

我们提出以端到端的方式联合优化度量深度和表面法线。这种优化主要旨在利用深度数据集中大量可用的标注知识来改进法线估计，特别是在室外场景中，深度数据集的标注数量明显多于法线数据集。在我们的实验中，我们从社区收集了来自 14 个室外数据集的带有深度标注的 9488 K 张图像，而室外带有法线标注的图像少于 20 K 张，如表 5 所示。

To facilitate knowledge flow across the depth and normal, we implement the learning-based optimization with recurrent refinement blocks, as depicted in Fig 10. Unlike previous monocular methods [42], [44], our method updates both depth and normal iteratively through these blocks. Inspired by RAFT [45], [46], we iteratively optimize the intermediate low-resolution depth $\widehat{\mathbf{D}}_c$ and unnormalized normal $\widehat{\mathbf{N}}_u$, where denotes low resolution prediction $\widehat{\mathbf{D}}_c \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4}}$, and $\widehat{\mathbf{N}}_u \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 3}$, and the subscript $_c$ means the depth $\widehat{\mathbf{D}}_c$ is in canonical space. As sketched in Fig. 10, $\widehat{\mathbf{D}}_c^t$ and $\widehat{\mathbf{N}}_u^t$ represent the low-resolution depth and normal optimized after step t , where $t = 0, 1, 2, \ldots, T$ denotes the step index. Initially, at step $t = 0$, $\widehat{\mathbf{D}}_c^0$ and $\widehat{\mathbf{N}}_u^0$ are given by the decoder. In addition to updating depth and normal, the optimization module also updates hidden feature maps $\mathbf{H}^t$, which are initialized by the decoder. During each iteration, the learned recurrent block $\mathcal{F}$ output updates $\Delta\widehat{\mathbf{D}}_c, \Delta\widehat{\mathbf{N}}_u$ and renews the hidden features $\mathbf{H}$:

为了促进深度和法线之间的知识流动，我们使用循环细化块实现基于学习的优化，如图 10 所示。与之前的单目方法 [42]、[44] 不同，我们的方法通过这些块迭代更新深度和法线。受 RAFT [45]、[46] 的启发，我们迭代优化中间低分辨率深度 $\widehat{\mathbf{D}}_c$ 和未归一化法线 $\widehat{\mathbf{N}}_u$ ，其中表 □ 示低分辨率预测 $\widehat{\mathbf{D}}_c \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4}}$ ，以及 $\widehat{\mathbf{N}}_u \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 3}$ ，下标 $_c$ 表示深度 $\widehat{\mathbf{D}}_c$ 处于规范空间。如图所示，$10, \widehat{\mathbf{D}}_c^t$ 和 $\widehat{\mathbf{N}}_u^t$ 表示在步骤 t 之后优化的低分辨率深度和法线，其中 $t = 0, 1, 2, \ldots, T$ 表示步骤索引。最初，在步骤 $t = 0, \widehat{\mathbf{D}}_c^0$ ，$\widehat{\mathbf{N}}_u^0$ 由解码器给出。除了更新深度和法线外，优化模块还更新隐藏特征图 $\mathbf{H}^t$ ，这些特征图由解码器初始化。在每次迭代中，学习到的循环块 $\mathcal{F}$ 的输出更新 $\Delta\widehat{\mathbf{D}}_c, \Delta\widehat{\mathbf{N}}_u$ 并更新隐藏特征 $\mathbf{H}$:

$$\Delta\widehat{\mathbf{D}}_c^{t+1}, \Delta\widehat{\mathbf{N}}_u^{t+1}, \mathbf{H}^{t+1} = \mathcal{F}\left(\widehat{\mathbf{D}}_u^t, \widehat{\mathbf{N}}_u^t, \mathbf{H}^t, \mathbf{H}^0\right), \tag{2}$$

The updates are then applied for updating the predictions:

然后将这些更新应用于更新预测:

$$\widehat{\mathbf{D}}_c^{t+1} = \widehat{\mathbf{D}}_c^t + \Delta\widehat{\mathbf{D}}_c^{t+1}, \widehat{\mathbf{N}}_u^{t+1} = \widehat{\mathbf{N}}_u^t + \Delta\widehat{\mathbf{N}}_u^{t+1}, \tag{3}$$

To be more specific, the recurrent block $\mathcal{F}$ comprises a ConvGRU sub-block and two projection heads. First, the ConvGRU sub-block updates the hidden features $\mathbf{H}^t$ taking all the variables as inputs. Subsequently, the two branched projection heads $\mathcal{G}_d$ and $\mathcal{G}_n$ estimate the updates $\Delta\widehat{\mathbf{D}}^{t+1}$ and $\Delta\widehat{\mathbf{N}}^{t+1}$ respectively. A more comprehensive representation of Eq. 2, therefore, can be written as:

更具体地说，循环块 $\mathcal{F}$ 由一个卷积门控循环单元 (ConvGRU) 子块和两个投影头组成。首先，ConvGRU 子块将所有变量作为输入来更新隐藏特征 $\mathbf{H}^t$ 。随后，两个分支投影头 $\mathcal{G}_d$ 和 $\mathcal{G}_n$ 分别估计更新 $\Delta\widehat{\mathbf{D}}^{t+1}$ 和 $\Delta\widehat{\mathbf{N}}^{t+1}$ 。因此，公式 2 的更全面表示可以写成:

$$\mathbf{H}^{t+1} = \text{ConvGRU}\left(\widehat{\mathbf{D}}^t, \widehat{\mathbf{N}}^t, \mathbf{H}^0, \mathbf{H}^t\right), \tag{4}$$

$$\Delta\widehat{\mathbf{D}}^{t+1} = \mathcal{G}_d\left(\mathbf{H}^{t+1}\right),\ \Delta\widehat{\mathbf{N}}^{t+1} = \mathcal{G}_n\left(\mathbf{H}^{t+1}\right).$$

For detailed structures of the refinement module $\mathcal{F}$, we recommend readers refer to supplementary materials.

关于细化模块 $\mathcal{F}$ 的详细结构，我们建议读者参考补充材料。

After $T+1$ iterative steps, we obtain the well-optimized low-resolution predictions $\widehat{\mathbf{D}}_c^{T+1}$ and $\widehat{\mathbf{N}}_u^{T+1}$. These predictions are then up-sampled and post-processed to generate the final depth $\mathbf{D}_c$ and surface normal $\mathbf{N}$:

经过 $T+1$ 次迭代步骤，我们得到了优化良好的低分辨率预测结果 $\widehat{\mathbf{D}}_c^{T+1}$ 和 $\widehat{\mathbf{N}}_u^{T+1}$ 。然后对这些预测结果进行上采样和后处理，以生成最终的深度图 $\mathbf{D}_c$ 和表面法线图 $\mathbf{N}$:

$$\mathbf{D}_c = \mathcal{H}_d\left(\text{upsample}\left(\widehat{\mathbf{D}}_c^{T+1}\right)\right) \tag{5}$$

$$\mathbf{N} = \mathcal{H}_n\left(\text{upsample}\left(\widehat{\mathbf{N}}_u^{T+1}\right)\right),$$

where $\mathcal{H}_d$ is the ReLU function to guarantee depth is nonnegative, and $\mathcal{H}_n$ represents normalization to ensure $\| \mathbf{n} \| = 1$ for all pixels.

其中 $\mathcal{H}_d$ 是 ReLU 函数，用于确保深度值为非负，$\mathcal{H}_n$ 表示归一化操作，以保证所有像素的 $\| \mathbf{n} \| =$ 为 1。

In a general formulation, the end-to-end network in Fig. 10 can be rewritten as:

一般来说，图 10 中的端到端网络可以重写为:

$$\mathbf{D}_c, \mathbf{N} = \mathcal{N}_{d-n}\left(\mathbf{I}_c, \theta\right) \tag{6}$$

where θ is the network’ s ($\mathcal{N}_{d-n}$) parameters.

其中 θ 是网络的 $(\mathcal{N}_{d-n})$ 参数。

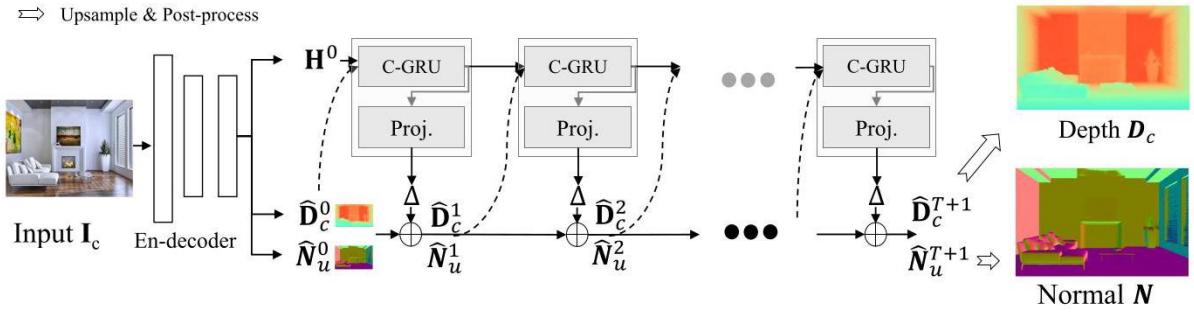

Fig. 10 - Joint depth and normal optimization. In the canonical space, we deploy recurrent blocks composed of ConvGRU sub-blocks (C-RGU) and projection heads (Proj.) to predict the updates Δ. During optimization, intermediate low-resolution depth and normal $\widehat{\mathbf{D}}_c^0\widehat{\mathbf{N}}_u^0$ are initially given by the decoder, and then iteratively refined by the predicted updates Δ. After $T+1$ iterations, the optimized intermediate predictions $\widehat{\mathbf{D}}_c^{T+1}\widehat{\mathbf{N}}_u^{T+1}$ are upsampled and post-processed to obtain the final depth $\mathbf{D}_c$ in the canonical space and the final normal $\mathbf{N}$.

图 10 - 联合深度和法线优化。在规范空间中，我们部署由卷积门控循环单元子块 (C - RGU) 和投影头 (Proj.) 组成的循环块来预测更新量 Δ。在优化过程中，中间低分辨率的深度图和法线图 $\widehat{\mathbf{D}}_c^0\widehat{\mathbf{N}}_u^0$ 最初由解码器给出，然后通过预测的更新量 Δ 进行迭代细化。经过 $T+1$ 次迭代后，优化后的中间预测结果 $\widehat{\mathbf{D}}_c^{T+1}\widehat{\mathbf{N}}_u^{T+1}$ 被上采样并进行后处理，以获得规范空间中的最终深度图 $\mathbf{D}_c$ 和最终法线图 $\mathbf{N}$。

3.4 Supervision

3.4 监督

The training objective is:

训练目标是:

$$\min_{\theta} L\left(\mathcal{N}_{d-n}\left(\mathbf{I}_c, \theta\right), \mathbf{D}_c^*, \mathbf{N}^*\right) \tag{7}$$

where $\mathbf{D}_c^*$ and $\mathbf{I}_c$ are transformed ground-truth depth labels and images in the canonical space $c, \mathbf{N}^*$ denotes normal labels, L is the supervision loss to be illustrated as following.

其中 $\mathbf{D}_c^*$ 和 $\mathbf{I}_c$ 是转换后的真实深度标签和规范空间中的图像，$c, \mathbf{N}^*$ 表示法线标签，L 是后续将说明的监督损失。

Random proposal normalization loss. To boost the performance of depth estimation, we propose a random proposal normalization loss (RPNL). The scale-shift invariant loss [25], [27] is widely applied for the affine-invariant depth estimation, which decouples the depth scale to emphasize the single image distribution. However, such normalization based on the whole image inevitably squeezes the fine-grained depth difference, particularly in close regions. Inspired by this, we propose to randomly crop several patches $\left(p_{i(i=0,\ldots,M)} \in \mathbb{R}^{h_i \times w_i}\right)$ from the ground

truth $\mathbf{D}_c^*$ and the predicted depth $\mathbf{D}_c$. Then we employ the median absolute deviation normalization [90] for paired patches. By normalizing the local statistics, we can enhance local contrast. The loss function is as follows:

随机提议归一化损失。为了提高深度估计的性能，我们提出了一种随机提议归一化损失 (RPNL)。尺度 - 平移不变损失 [25]、[27] 广泛应用于仿射不变深度估计，它将深度尺度解耦以强调单图像分布。然而，这种基于整幅图像的归一化不可避免地会压缩细粒度的深度差异，特别是在近距离区域。受此启发，我们提议从真实标签 $\mathbf{D}_c^*$ 和预测深度 $\mathbf{D}_c$ 中随机裁剪出几个图像块 $\left(p_{i(i=0,\ldots,M)} \in \mathbb{R}^{h_i \times w_i}\right)$ 。然后我们对成对的图像块采用中位数绝对偏差归一化 [90]。通过对局部统计量进行归一化，我们可以增强局部对比度。损失函数如下:

$$L_{\text{RPNL}} = \frac{1}{MN} \sum_{p_i}^{M} \sum_{j}^{N} \left| \frac{d^*_{p_i,j} - \mu\left(d^*_{p_i,j}\right)}{\frac{1}{N} \sum_{j}^{N} \left| d^*_{p_i,j} - \mu\left(d^*_{p_i,j}\right) \right|} - \frac{d_{p_i,j} - \mu\left(d_{p_i,j}\right)}{\frac{1}{N} \sum_{j}^{N} \left| d_{p_i,j} - \mu\left(d_{p_i,j}\right) \right|} \right| \tag{8}$$

where $d^* \in \mathbf{D}_c^*$ and $d \in \mathbf{D}_c$ are the ground truth and predicted depth respectively. $\mu(\cdot)$ and is the median of depth. M is the number of proposal crops, which is set to 32. During training, proposals are randomly cropped from the image by 0.125 to 0.5 of the original size. Furthermore, several other losses are employed, including the scale-invariant logarithmic loss [60] L_{silog} , pair-wise normal regression loss [25] L_{PWN} , virtual normal loss [18] L_{VNL} . Note L_{silog} is a variant of L1 loss. The overall losses are as follows.

其中 $d^* \in \mathbf{D}_c^*$ 和 $d \in \mathbf{D}_c$ 分别是真实标签和预测深度。$\mu(\cdot)$ 是深度的中位数。M 是提议裁剪的数量，设置为 32。在训练过程中，提议从图像中随机裁剪，裁剪大小为原始大小的 0.125 到 0.5。此外，还采用了其他几种损失，包括尺度不变对数损失 [60] L_{silog} 、成对法线回归损失 [25] L_{PWN} 、虚拟法线损失 [18] L_{VNL} 。注意 L_{silog} 是 L1 损失的一种变体。总体损失如下。

$$L_d = L_{\text{PWN}} + L_{\text{VNL}} + L_{\text{silog}} + L_{\text{RPNL}}. \tag{9}$$

Normal loss. To supervise normal prediction, we employ two distinct loss functions depending on the availability of ground-truth (GT) normals $\mathbf{N}^*$. As presented in Fig. 9, when GT normals are provided, we utilize an aleatoric uncertainty-aware loss [33] $(L_n(\cdot))$ to supervise prediction N. Alternatively, in the absence of GT normals, we propose a consistency loss $L_{d-n}(\mathbf{D}, \mathbf{N})$ to align the predicted depth and normal. This loss is computed based on the similarity between a pseudo-normal map generated from the predicted depth using the least square method [41], and the predicted normal itself. Different from previous methods, [33], [41], this loss operates as a self-supervision mechanism, requiring no depth or normal ground truth labels. Note that here we use the depth $\mathbf{D}$ in the real world instead of the one $\mathbf{D}_c$ in the canonical space to calculate depth-normal consistency. The overall losses are as follows.

正常损失。为了监督法线预测，我们根据真实法线 (GT) $\mathbf{N}^*$ 的可用性采用两种不同的损失函数。如图 9 所示，当提供真实法线时，我们利用一种考虑随机不确定性的损失 [33] $(L_n(\cdot))$ 来监督预测法线 N。或者，在没有真实法线的情况下，我们提出一种一致性损失 $L_{d-n}(\mathbf{D}, \mathbf{N})$ 来对齐预测的深度和法线。该损失是基于使用最小二乘法 [41] 从预测深度生成的伪法线图与预测法线本身之间的相似度来计算的。与之前的方法 [33]、[41] 不同，这种损失作为一种自监督机制运行，不需要深度或法线的真实标签。请注意，这里我们使用真实世界中的深度 **D** 而不是规范空间中的深度 $\mathbf{D}_c$ 来计算深度 - 法线一致性。总体损失如下。

$$L = w_d L_d(\mathbf{D}_c, \mathbf{D}_c^*) + w_n L_n(\mathbf{N}, \mathbf{N}^*) + w_{d-n} L_{d-n}(\mathbf{N}, \mathbf{I} \tag{10}$$

, where $w_d = 0.5, w_n = 1, w_{d-n} = 0.01$ serve as weights to balance the loss items.

其中 $w_d = 0.5, w_n = 1, w_{d-n} = 0.01$ 作为权重来平衡损失项。

4 EXPERIMENTS

4 实验

Dataset details. We have meticulously assembled a comprehensive dataset incorporating 16 publicly available RGB-D datasets, comprising a cumulative total of over 16 million data points specifically intended for training purposes. This dataset encompasses a diverse array of both indoor and outdoor scenes. Notably, approximately 10 million frames within the dataset are annotated with normals, with a predominant focus on annotations relating to indoor scenes. It is noteworthy to highlight that all datasets have provided camera intrinsic parameters. Additionally, beyond the test split of training datasets, we have procured 7 previously unobserved datasets to facilitate robustness and generalization evaluations. Detailed descriptions of the utilized training and testing data are provided in Table 5.

数据集详情。我们精心整理了一个综合数据集，其中包含 16 个公开可用的 RGB - D 数据集，总共包含超过 1600 万个专门用于训练的数据点。该数据集涵盖了各种室内和室外场景。值得注意的是，数据集中约 1000 万帧标注了法线，主要集中在室内场景的标注上。值得强调的是，所有数据集都提供了相机内参。此外，除了训练数据集的测试分割外，我们还获取了 7 个之前未见过的数据集，以进行鲁棒性和泛化性评估。表 5 提供了所使用的训练和测试数据的详细描述。

Implementation details. In our experiments, we employ different network architectures and aim to provide diverse choices for the community, including convnets and transformers. For convnets, we employ an UNet architecture with the ConvNext-large [102] backbone. ImageNet-22K pre-trained weights are used for initialization. For transformers, we apply DINO v2-reg [85], [103] vision transformers [86] (ViT) as backbones, DPT [28] as decoders.

实现细节。在我们的实验中，我们采用不同的网络架构，旨在为社区提供多样化的选择，包括卷积网络 (convnets) 和 Transformer 网络。对于卷积网络，我们采用具有 ConvNext - large [102] 主干的 UNet 架构。使用在 ImageNet - 22K 上预训练的权重进行初始化。对于 Transformer 网络，我们应用 DINO v2 - reg [85]、[103] 视觉 Transformer(ViT)[86] 作为主干，DPT [28] 作为解码器。

We use AdamW with a batch size of 192, an initial learning rate 0.0001 for all layers, and the polynomial decaying method with the power of 0.9 . We train our models on 48 A100 GPUs for 800k iterations. Following the DiverseDepth [18], we balance all datasets in a mini-batch to ensure each dataset accounts for an almost equal ratio. During training, images are processed by the canonical camera transformation module, flipped horizontally with a 50% chance, and then randomly cropped into 512×960 pixels for convnets and 616×1064 for vision transformers. In the ablation experiments, training settings are different as we sample 5000 images from each dataset for training. We trained on 8GPUs for 150 K iterations. Details of networks architectures, training setups, and efficiency analysis are presented in the supplementary materials. Fine-tuning experiments on KITTI and NYU are conducted on 8 GPUs with 20 K further steps.

我们使用 AdamW 优化器，批量大小为 192，所有层的初始学习率为 0.0001，并采用幂为 0.9 的多项式衰减方法。我们在 48 个 A100 GPU 上对我们的模型进行 800k 次迭代训练。遵循 DiverseDepth [18] 的方法，我们在一个小批量中平衡所有数据集，以确保每个数据集所占比例几乎相等。在训练期间，图像由规范相机变换模块进行处理，有 50% 的概率进行水平翻转，然后随机裁剪为卷积网络的 512×960 像素和视觉 Transformer 的 616×1064 像素。在消融实验中，训练设置不同，因为我们从每个数据集中采样 5000 张图像进行训练。我们在 8GPUs 上进行 150 K 次迭代训练。网络架构、训练设置和效率分析的详细信息在补充材料中给出。在 KITTI 和 NYU 上的微调实验在 8 个 GPU 上进行 20 K 步。

TABLE 1 - Quantitative comparison on NYUv2 and KITTI metric depth benchmarks. Methods overfitting the benchmark are marked with grey, while robust depth estimation methods are in blue. 'ZS' denotes the zero-shot testing, and 'FT' means the method is further finetuned on the benchmark. Among all zero-shot testing (ZS) results, our methods performs the best and is even better than overfitting methods. Further fine-tuning (FT) helps our method surpass all known methods, ranked by the averaged ranking among all metrics. Best results are in bold and second bests are underlined.

表 1 - 在 NYUv2 和 KITTI 深度度量基准上的定量比较。过拟合基准的方法用灰色标记，而鲁棒的深度估计方法用蓝色标记。'ZS' 表示零样本测试，'FT' 表示该方法在基准上进一步微调。在所有零样本测试 (ZS) 结果中，我们的方法表现最佳，甚至优于过拟合方法。进一步微调 (FT) 帮助我们的方法在所有指标的平均排名中超过所有已知方法。最佳结果用粗体表示，第二好的结果用下划线表示。

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	AbsRel↓	log 10 ↓	RMS↓
NYUv2 Metric Depth Benchmark						
Li et al.. [91]	0.788	0.958	0.991	0.143	0.063	0.635
Laina et al.. [92]	0.811	0.953	0.988	0.127	0.055	0.573
VNL [30]	0.875	0.976	0.994	0.108	0.048	0.416
TrDepth [20]	0.900	0.983	0.996	0.106	0.045	0.365
Adabins [19]	0.903	0.984	0.997	0.103	0.044	0.364
NeWCRFs [17]	0.922	0.992	0.998	0.095	0.041	0.334
IEBins [44]	0.936	0.992	0.998	0.087	0.038	0.314
ZeroDepth [61] ZS	0.901	0.961	-	0.100	-	0.380
Polymax [34] ZS	0.969	0.996	0.999	0.067	0.029	0.250
ZoeDepth [36] FT	0.953	0.995	0.999	0.077	0.033	0.277
ZeroDepth [61] FT	0.954	0.995	1.000	0.074	0.103	0.269
DepthAnything [93] FT	0.984	0.998	1.000	0.056	0.024	0.206
Ours Conv-L CSTM image ZS	0.925	0.983	0.994	0.092	0.040	0.341
Ours Conv-L CSTM label ZS	0.944	0.986	0.995	0.083	0.035	0.310
Ours ViT-L CSTM label ZS	0.975	0.994	0.998	0.063	0.028	0.251
Ours ViT-g CSTM label ZS	0.980	0.997	0.999	0.067	0.030	0.260
Ours ViT-L CSTM label FT	0.989	0.998	1.000	0.047	0.020	0.183
Ours ViT-g CSTM_label FT	0.987	0.997	0.999	0.045	0.015	0.187

方法	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	绝对相对误差 (AbsRel)↓	log 10 ↓	均方根误差 (RMS)↓
NYUv2 深度指标基准测试						
李等人 [91]	0.788	0.958	0.991	0.143	0.063	0.635
莱娜等人 [92]	0.811	0.953	0.988	0.127	0.055	0.573
视觉法线学习 (VNL) [30]	0.875	0.976	0.994	0.108	0.048	0.416
Transformer 深度估计 (TrDepth) [20]	0.900	0.983	0.996	0.106	0.045	0.365
自适应分箱 (Adabins) [19]	0.903	0.984	0.997	0.103	0.044	0.364
新型加权条件随机场 (NeWCRFs) [17]	0.922	0.992	0.998	0.095	0.041	0.334
信息熵分箱 (IEBins) [44]	0.936	0.992	0.998	0.087	0.038	0.314
零样本深度估计 (ZeroDepth) [61] 零样本 (ZS)	0.901	0.961	-	0.100	-	0.380
多峰深度估计 (Polymax) [34] 零样本 (ZS)	0.969	0.996	0.999	0.067	0.029	0.250
佐伊深度估计 (ZoeDepth) [36] 微调 (FT)	0.953	0.995	0.999	0.077	0.033	0.277
零样本深度估计 (ZeroDepth) [61] 微调 (FT)	0.954	0.995	1.000	0.074	0.103	0.269
任意深度估计 (DepthAnything) [93] 微调 (FT)	0.984	0.998	1.000	0.056	0.024	0.206
我们的方法: 卷积大模型 (Conv - L) 条件空间变换器模块 (CSTM) 图像零样本 (ZS)	0.925	0.983	0.994	0.092	0.040	0.341
我们的方法: 卷积大模型 (Conv - L) 条件空间变换器模块 (CSTM) 标签零样本 (ZS)	0.944	0.986	0.995	0.083	0.035	0.310
我们的方法: 视觉 Transformer 大模型 (ViT - L) 条件空间变换器模块 (CSTM) 标签零样本 (ZS)	0.975	0.994	0.998	0.063	0.028	0.251
我们的方法: 视觉 Transformer 巨型模型 (ViT - g) 条件空间变换器模块 (CSTM) 标签零样本 (ZS)	0.980	0.997	0.999	0.067	0.030	0.260
我们的方法: 视觉 Transformer 大模型 (ViT - L) 条件空间变换器模块 (CSTM) 标签微调 (FT)	0.989	0.998	1.000	0.047	0.020	0.183
我们的方法: 视觉 Transformer 巨型模型 (ViT - g) 条件空间变换器模块 (CSTM) 标签微调 (FT)	0.987	0.997	0.999	0.045	0.015	0.187

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	AbsRel $\downarrow$	RMS $\downarrow$	RMS_log $\downarrow$
KITTI Metric Depth Benchmark						
Guo et al. [94]	0.902	0.969	0.986	0.090	3.258	0.168
VNL [30]	0.938	0.990	0.998	0.072	3.258	0.117
TrDepth [20]	0.956	0.994	0.999	0.064	2.755	0.098
Adabins [19]	0.964	0.995	0.999	0.058	2.360	0.088
NeWCRFs [17]	0.974	0.997	0.999	0.052	2.129	0.079
IEBins [44]	0.978	0.998	0.999	0.050	2.011	0.075
ZeroDepth [61] ZS	0.910	0.980	0.996	0.102	4.044	0.172
ZoeDepth [36] FT	0.971	0.996	0.999	0.057	2.281	0.082
ZeroDepth [61] FT	0.968	0.995	0.999	0.053	2.087	0.083
DepthAnything [93] FT	0.982	0.998	1.000	0.046	1.869	0.069
Ours Conv-L CSTM_image ZS	0.967	0.995	0.999	0.060	2.843	0.087
Ours Conv-L CSTM label ZS	0.964	0.993	0.998	0.058	2.770	0.092
Ours ViT-L CSTM label ZS	0.974	0.995	0.999	0.052	2.511	0.074
Ours ViT-g CSTM label ZS	0.977	0.996	0.999	0.051	2.403	0.080
Ours ViT-L CSTM_label FT	0.985	0.998	0.999	0.044	1.985	0.064
Ours ViT-g CSTM_label FT	0.989	0.998	1.000	0.039	1.766	0.060

方法	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	绝对相对误差 (AbsRel)↓	均方根误差 (RMS) ↓	对数均方根误差 (RMS_log) ↓
基蒂数据集 (KITTI) 深度评估基准						
郭等人 [94]	0.902	0.969	0.986	0.090	3.258	0.168
视觉法线学习 (VNL) [30]	0.938	0.990	0.998	0.072	3.258	0.117
Transformer 深度估计 (TrDepth) [20]	0.956	0.994	0.999	0.064	2.755	0.098
自适应分箱 (Adabins) [19]	0.964	0.995	0.999	0.058	2.360	0.088
新型加权条件随机场 (NeWCRFs) [17]	0.974	0.997	0.999	0.052	2.129	0.079
改进的分箱 (IEBins) [44]	0.978	0.998	0.999	0.050	2.011	0.075
零样本深度估计 (ZeroDepth) [61] 零样本 (ZS)	0.910	0.980	0.996	0.102	4.044	0.172
佐伊深度估计 (ZoeDepth) [36] 微调 (FT)	0.971	0.996	0.999	0.057	2.281	0.082
零样本深度估计 (ZeroDepth) [61] 微调 (FT)	0.968	0.995	0.999	0.053	2.087	0.083
任意深度估计 (DepthAnything) [93] 微调 (FT)	0.982	0.998	1.000	0.046	1.869	0.069
我们的方法卷积大模型 (Conv - L) 基于图像的条件空间变换器模块 (CSTM_image) 零样本 (ZS)	0.967	0.995	0.999	0.060	2.843	0.087
我们的方法卷积大模型 (Conv - L) 基于标签的条件空间变换器模块 (CSTM label) 零样本 (ZS)	0.964	0.993	0.998	0.058	2.770	0.092
我们的方法视觉 Transformer 大模型 (ViT - L) 基于标签的条件空间变换器模块 (CSTM label) 零样本 (ZS)	0.974	0.995	0.999	0.052	2.511	0.074
我们的方法视觉 Transformer 巨型模型 (ViT - g) 基于标签的条件空间变换器模块 (CSTM label) 零样本 (ZS)	0.977	0.996	0.999	0.051	2.403	0.080
我们的方法视觉 Transformer 大模型 (ViT - L) 基于标签的条件空间变换器模块 (CSTM_label) 微调 (FT)	0.985	0.998	0.999	0.044	1.985	0.064
我们的方法视觉 Transformer 巨型模型 (ViT - g) 基于标签的条件空间变换器模块 (CSTM_label) 微调 (FT)	0.989	0.998	1.000	0.039	1.766	0.060

Evaluation details for monocular depth and normal estimation. a) To demonstrate the robustness of our metric depth estimation method, we evaluate on 7 zero-shot benchmarks, including NYUv2, KITTI [106], ScanNet [107], NuScenes [108], iBIMS-1 [109], and DIODE [110] (both indoor and outdoor). Following previous studies, we use metrics such as absolute relative error (AbsRel), accuracy under threshold $\left(\delta_i < 1.25^i, i = 1, 2, 3\right)$, root mean squared error (RMS), root mean squared error in log space (RMS_log), and log 10 error (log 10). We report results for zero-shot and fine-tuning testing on the KITTI and NYU benchmarks. b) For normal estimation tasks and ablations, employ several error metrics to assess performance. Specifically, we calculate the mean (mean), median (median), and rooted mean square (RMS_normal) of the angular error as well as the accuracy under threshold of $\{11.25^\circ, 22.5^\circ, 30.0^\circ\}$ consistent with methodologies established in previous studies [33]. We conduct in-domain evaluation using the Scannet dataset, while the NYU and iBIMS-1 datasets are reserved for zero-shot generalization testing. c) Furthermore, we also follow current affine-invariant depth benchmarks [25], [29] (Tab. 4) to evaluate the generalization ability on 5 zero-shot datasets, i.e., NYUv2, DIODE, ETH3D, ScanNet [107], and KITTI. We mainly compare with large-scale data trained models. Note that in this benchmark we follow existing methods to apply the scale shift alignment before evaluation.

单目深度和法线估计的评估细节。a) 为了证明我们的度量深度估计方法的鲁棒性，我们在 7 个零样本基准数据集上进行评估，包括 NYUv2、KITTI [106]、ScanNet [107]、NuScenes [108]、iBIMS - 1 [109] 和 DIODE [110](包括室内和室外)。遵循先前的研究，我们使用诸如绝对相对误差 (AbsRel)、阈值 $\left(\delta_i < 1.25^i, i = 1, 2, 3\right)$ 下的准确率、均方根误差 (RMS)、对数空间中的均方根误差 (RMS_log) 和 $\log 10$ 误差 ($\log 10$) 等指标。我们报告了在 KITTI 和 NYU 基准数据集上的零样本和微调测试结果。b) 对于法线估计任务和消融实验，采用了几种误差指标来评估性能。具体而言，我们计算角度误差的均值 (mean)、中位数 (median) 和均方根 (RMS_normal)，以及与先前研究 [33] 中确立的方法一致的阈值 $\{11.25°, 22.5°, 30.0°\}$ 下的准确率。我们使用 Scannet 数据集进行域内评估，而 NYU 和 iBIMS - 1 数据集则用于零样本泛化测试。c) 此外，我们还遵循当前的仿射不变深度基准 [25]、[29](表 4)，在 5 个零样本数据集上评估泛化能力，即 NYUv2、DIODE、ETH3D、ScanNet [107] 和 KITTI。我们主要与大规模数据训练的模型进行比较。请注意，在这个基准测试中，我们遵循现有方法在评估前进行尺度偏移对齐。

TABLE 2 - Quantitative comparison of surface normals on NYUv2, ibims-1, and ScanNet normal benchmarks. 'ZS' means zero-shot testing and 'FT' performs post fine-tuneing on the target dataset. Methods trained only on NYU are highlighted with grey. Best results are in bold and second bests are underlined. Our method ranks first over all benchmarks.

表 2 - 在 NYUv2、ibims - 1 和 ScanNet 法线基准数据集上的表面法线定量比较。'ZS' 表示零样本测试，'FT' 表示在目标数据集上进行后微调。仅在 NYU 上训练的方法用灰色突出显示。最佳结果用粗体表示，次佳结果用下划线表示。我们的方法在所有基准测试中排名第一。

Method	**11.25°** ↑	22.5° ↑	30° ↑	mean↓	median↓	RMS_normal ↓
NYUv2 Normal Benchmar						
Ladicky et al. [69]	0.275	0.490	0.587	33.5	23.1	-
Fouhey et al.. [95]	0.405	0.541	0.589	35.2	17.9	-
Deep3D [66]	0.420	0.612	0.682	20.9	13.2	-
Eigen et al. [67]	0.444	0.672	0.759	20.9	13.2	-
SkipNet [96]	0.479	0.700	0.778	19.8	12.0	28.2
SURGE [97]	0.473	0.689	0.766	20.6	12.2	-
GeoNet [41]	0.484	0.715	0.795	19.0	11.8	26.9
PAP [98]	0.488	0.722	0.798	18.6	11.7	25.5
GeoNet++ [65]	0.502	0.732	0.807	18.5	11.2	26.7
Bae et al. [33]	0.622	0.793	0.852	14.9	7.5	23.5
FrameNet [40] ZS	0.507	0.720	0.795	18.6	11.0	26.8
VPLNet [99] ZS	0.543	0.738	0.807	18.0	9.8	-
TiltedSN [32] ZS	0.598	0.774	0.834	16.1	8.1	25.1
Omnidata [35] ZS	0.577	0.777	0.838	16.7	9.6	25.0
Bae et al. [33] ZS	0.597	0.775	0.837	16.0	8.4	24.7
Polymax [34] ZS	0.656	0.822	0.878	13.1	7.1	20.4
Ours ViT-L CSTM label ZS	0.662	0.831	0.881	13.1	7.1	21.1
Ours ViT-g CSTM_label ZS	0.664	0.831	0.881	13.3	7.0	21.3
Ours ViT-L CSTM_label FT	0.688	0.849	0.898	12.0	6.5	19.2
Ours ViT-g CSTM_label FT	0.662	0.837	0.889	13.2	7.5	20.2

方法	**11.25°** ↑	22.5° ↑	30° ↑	均值 ↓	中位数 ↓	均方根法线 ↓
NYUv2 法线基准测试						
拉迪茨基等人 [69]	0.275	0.490	0.587	33.5	23.1	-
福伊等人 [95]	0.405	0.541	0.589	35.2	17.9	-
深度 3D [66]	0.420	0.612	0.682	20.9	13.2	-
艾根等人 [67]	0.444	0.672	0.759	20.9	13.2	-
跳跃网络 [96]	0.479	0.700	0.778	19.8	12.0	28.2
浪涌 [97]	0.473	0.689	0.766	20.6	12.2	-
地理网络 [41]	0.484	0.715	0.795	19.0	11.8	26.9
PAP [98]	0.488	0.722	0.798	18.6	11.7	25.5
地理网络 ++ [65]	0.502	0.732	0.807	18.5	11.2	26.7
裴等人 [33]	0.622	0.793	0.852	14.9	7.5	23.5
帧网络 [40] 零样本	0.507	0.720	0.795	18.6	11.0	26.8
VPL 网络 [99] 零样本	0.543	0.738	0.807	18.0	9.8	-
倾斜 SN [32] 零样本	0.598	0.774	0.834	16.1	8.1	25.1
全数据 [35] 零样本	0.577	0.777	0.838	16.7	9.6	25.0
裴等人 [33] 零样本	0.597	0.775	0.837	16.0	8.4	24.7
多极大 [34] 零样本	0.656	0.822	0.878	13.1	7.1	20.4
我们的 ViT - L CSTM 标签零样本	0.662	0.831	0.881	13.1	7.1	21.1
我们的 ViT - g CSTM 标签零样本	0.664	0.831	0.881	13.3	7.0	21.3
我们的 ViT - L CSTM 标签微调	0.688	0.849	0.898	12.0	6.5	19.2
我们的 ViT - g CSTM 标签微调	0.662	0.837	0.889	13.2	7.5	20.2

ibims-1 Normal Benchmark						
VNL [30] ZS	0.179	0.386	0.494	39.8	30.4	51.0
BTS [100] ZS	0.130	0.295	0.400	44.0	37.8	53.5
Adabins [19] ZS	0.180	0.387	0.506	37.1	29.6	46.9
IronDepth [42] ZS	0.431	0.639	0.716	25.3	14.2	37.4
Omnidata [101] ZS	0.647	0.734	0.768	20.8	7.7	35.1
Ours ViT-L CSTM_label ZS	0.694	0.758	0.785	19.4	5.7	34.9
Ours ViT-g CSTM_label ZS	0.697	0.762	0.788	19.6	5.7	35.2

ibims - 1 标准基准 (ibims - 1 Normal Benchmark)						
VNL [30] 零样本 (VNL [30] ZS)	0.179	0.386	0.494	39.8	30.4	51.0
BTS [100] 零样本 (BTS [100] ZS)	0.130	0.295	0.400	44.0	37.8	53.5
Adabins [19] 零样本 (Adabins [19] ZS)	0.180	0.387	0.506	37.1	29.6	46.9
IronDepth [42] 零样本 (IronDepth [42] ZS)	0.431	0.639	0.716	25.3	14.2	37.4
Omnidata [101] 零样本 (Omnidata [101] ZS)	0.647	0.734	0.768	20.8	7.7	35.1
我们的 ViT - L CSTM 标签零样本 (Ours ViT - L CSTM_label ZS)	0.694	0.758	0.785	19.4	5.7	34.9
我们的 ViT - g CSTM 标签零样本 (Ours ViT - g CSTM_label ZS)	0.697	0.762	0.788	19.6	5.7	35.2

ScanNet Normal Benchmark						
Omnidata [101]	0.629	0.806	0.847	15.1	8.6	23.1
FrameNet [40]	0.625	0.801	0.858	14.7	7.7	22.8
VPLNet [99]	0.663	0.818	0.870	12.6	6.0	21.1
TiltedSN [32]	0.693	0.839	0.886	12.6	6.0	21.1
Bae et al. [33]	0.711	0.854	0.898	11.8	5.7	20.0
Ours ViT-L CSTM_label	0.760	0.885	0.923	9.9	5.3	16.4
Ours ViT-g CSTM_label	0.778	0.901	0.935	9.2	5.0	15.3

扫描网络法线基准测试 (ScanNet Normal Benchmark)						
全数据 (Omnidata) [101]	0.629	0.806	0.847	15.1	8.6	23.1
帧网络 (FrameNet) [40]	0.625	0.801	0.858	14.7	7.7	22.8
VPL 网络 (VPLNet) [99]	0.663	0.818	0.870	12.6	6.0	21.1
倾斜扫描网络 (TiltedSN) [32]	0.693	0.839	0.886	12.6	6.0	21.1
贝等人 (Bae et al.) [33]	0.711	0.854	0.898	11.8	5.7	20.0
我们的 ViT - L CSTM 标签	0.760	0.885	0.923	9.9	5.3	16.4
我们的 ViT - g CSTM 标签	0.778	0.901	0.935	9.2	5.0	15.3

We report results with different canonical transformation methods (CSTM_lable and CSTM_image) on the ConvNext-Large model (Conv-L in Tab. 1 and Tab. 2). As CSTM_label is slightly better, more results using this method from multi-size ViT-models (ViT-S for Small, ViT-L for Large, ViT-g for giant2) are reported. Note that all models for zero-shot testing use the same checkpoints except for fine-tuning experiments.

我们报告了在 ConvNext-Large 模型 (表 1 和表 2 中的 Conv-L) 上使用不同规范变换方法 (规范变换方法标签版 (CSTM_lable) 和规范变换方法图像版 (CSTM_image)) 的结果。由于规范变换方法标签版 (CSTM_label) 略胜一筹，因此报告了更多使用该方法在多尺寸视觉 Transformer 模型 (小尺寸的 ViT-S、大尺寸的 ViT-L、巨型的 ViT-g) 上的结果。请注意，除了微调实验外，所有用于零样本测试的模型都使用相同的检查点。

Evaluation details for reconstruction and SLAM. a) To evaluate our metric 3D reconstruction quality, we randomly sample 9 unseen scenes from NYUv2 and use colmap [111] to obtain the camera poses for multi-frame reconstruction. Chamfer l_1 distance and the F-score [112] are used to evaluate the reconstruction accuracy. b) In dense-SLAM experiments, following Li et al. [113], we test on the KITTI odometry benchmark [59] and evaluate the average translational RMS $(\%, t_{rel})$ and rotational RMS $(^\circ/100m, r_{rel})$ errors [59]. Evaluation on metric depth benchmarks. To evaluate the accuracy of predicted metric depth, firstly, we compare with state-of-the-art (SoTA) metric depth prediction methods on NYUv2 [58], KITTI [106]. We use the same model to do all evaluations. Results are reported in Tab. 1. Firstly, comparing with existing overfitting methods, which are trained on benchmarks for hundreds of epochs, our zero-shot testing ('ZS' in the table) without any fine-tuning or metric adjustment already achieves comparable or even better performance on some metrics. Then comparing with robust monocular depth estimation methods, such as Ze-rodepth [61] and ZoeDepth [36], our zero-shot testing is also better than them. Further post finetuning ('FT in the table') lifts our method to the 1st rank.

重建和同步定位与地图构建 (SLAM) 的评估细节。a) 为了评估我们的指标 3D 重建质量，我们从 NYUv2 数据集中随机采样 9 个未见场景，并使用 colmap [111] 获取多帧重建的相机位姿。使用倒角 l_1 距离和 F 分数 [112] 来评估重建精度。b) 在稠密同步定位与地图构建 (dense-SLAM) 实验中，遵循 Li 等人 [113] 的方法，我们在 KITTI 里程计基准测试 [59] 上进行测试，并评估平均平移均方根 $(\%, t_{rel})$ 误差和旋转均方根 $(^\circ/100m, r_{rel})$ 误差 [59]。度量深度基准测试的评估。为了评估预测度量深度的准确性，首先，我们在 NYUv2 [58]、KITTI [106] 数据集上与最先进 (SoTA) 的度量深度预测方法进行比较。我们使用相同的模型进行所有评估。结果报告在表 1 中。首先，与现有的过拟合方法 (这些方法在基准数据集上训练了数百个轮次) 相比，我们的零样本测试 (表中的'ZS') 在不进行任何微调或度量调整的情况下，在某些指标上已经取得了相当甚至更好的性能。然后，与鲁棒的单目深度估计方法 (如 ZeroDepth [61] 和 ZoeDepth [36]) 相比，我们的零样本测试也优于它们。进一步的后期微调 (表中的'FT') 使我们的方法排名第一。

TABLE 3 - Quantitative comparison with SoTA metric depth methods on 5 unseen benchmarks. For SoTA methods, we use their NYUv2 and KITTI models for indoor and outdoor scene evaluation respectively, while we use the same model for all zero-shot testing.

表 3 - 在 5 个未见基准数据集上与最先进 (SoTA) 度量深度方法的定量比较。对于最先进 (SoTA) 的方法，我们分别使用它们的 NYUv2 和 KITTI 模型进行室内和室外场景评估，而我们使用相同的模型进行所有零样本测试。

Method	Metric Head	DIODE(Indoor)	iBIMS-1 Indoor scenes (AbsRel↓/RMS↓)	DIODE(Outdoor)	ETH3D Outdoor scenes (AbsRel↓/RMS↓)	NuScenes
Adabins [19]	KITTI or NYU †	0.443 / 1.963	0.212 / 0.901	0.865 / 10.35	1.271 / 6.178	0.445 / 10.658
NewCRFs [17]	KITTI or NYU †	0.404 / 1.867	0.206 / 0.861	0.854 / 9.228	0.890 / 5.011	0.400 / 12.139
ZoeDepth [36]	KITTI and NYU ‡	0.400 / 1.581	0.169 / 0.711	0.269 / 6.898	0.545 / 3.112	0.504 / 7.717
Ours Conv-L CSTM_label	Unified	0.252 / 1.440	0.160 / 0.521	0.414 / 6.934	0.416 / 3.017	0.154 / 7.097
Ours Conv-L CSTM_image	Unified	0.268 / 1.429	0.144 / 0.646	0.535 / 6.507	0.342 / 2.965	0.147 / 5.889
Ours ViT-L CSTM_image	Unified	0.093 / 0.389	0.185 / 0.592	0.221 / 3.897	0.357 / 2.980	0.165 / 9.001
Ours ViT-g CSTM_image	Unified	0.081 / 0.359	0.249 / 0.611	0.201 / 3.671	0.363 / 2.999	0.129 / 6.993

方法	度量头	DIODE(室内)iBIMS - 1 室内场景 (绝对相对误差 ↓/均方根误差 ↓)		DIODE(室外)	ETH3D 室外场景 (绝对相对误差 ↓/均方根误差 ↓)	NuScenes 数据集
自适应分箱法 (Adabins) [19]	基带数据集 (KITTI) 或组约大学数据集 (NYU) †	0.443 / 1.963	0.212 / 0.901	0.865 / 10.35	1.271 / 6.178	0.445 / 10.658
新型条件随机场 (NewCRFs) [17]	基带数据集 (KITTI) 或组约大学数据集 (NYU) †	0.404 / 1.867	0.206 / 0.861	0.854 / 9.228	0.890 / 5.011	0.400 / 12.139
佐伊深度估计法 (ZoeDepth) [36]	基带数据集 (KITTI) 和组约大学数据集 (NYU) ‡	0.400 / 1.581	0.169 / 0.711	0.269 / 6.898	0.545 / 3.112	0.504 / 7.717
我们的卷积 - 大模型 (Conv - L) 条件空间转换模块标签 (CSTM_label)	统一的	0.252 / 1.440	0.160 / 0.521	0.414 / 6.934	0.416 / 3.017	0.154 / 7.097
我们的卷积 - 大模型 (Conv - L) 条件空间转换模块图像 (CSTM_image)	统一的	0.268 / 1.429	0.144 / 0.646	0.535 / 6.507	0.342 / 2.965	0.147 / 5.889
我们的视觉 Transformer - 大模型 (ViT - L) 条件空间转换模块图像 (CSTM_image)	统一的	0.093 / 0.389	0.185 / 0.592	0.221 / 3.897	0.357 / 2.980	0.165 / 9.001
我们的视觉 Transformer - g 模型 (ViT - g) 条件空间转换模块图像 (CSTM_image)	统一的	0.081 / 0.359	0.249 / 0.611	0.201 / 3.671	0.363 / 2.999	0.129 / 6.993

† : Two different metric heads are trained on KITTI and NYU respectively. ‡ : Both metric heads are ensembled by an additional router.

† : 分别在 KITTI 和 NYU 数据集上训练了两个不同的度量头 (metric head)。‡ : 两个度量头通过一个额外的路由模块 (router) 进行集成。

TABLE 4 - Comparison with SoTA affine-invariant depth methods on 5 zero-shot transfer benchmarks. Our model significantly outperforms previous methods and sets new state-of-the-art. Following the benchmark setting, all methods have manually aligned the scale and shift.

表 4 - 在 5 个零样本迁移基准测试中与最先进的仿射不变深度方法的比较。我们的模型显著优于先前的方法，并创造了新的最先进水平。按照基准测试的设置，所有方法都手动对齐了尺度和偏移。

MethodBackbone	#Params	#Data		NYUv2		KITTI		DIODE(Full)		ScanNet		ETH3D	
		Pretrain	Train	AbsRel↓ δ_1 ↑		AbsRel↓ δ_1 ↑		AbsRel ↓ δ_1 ↑		AbsRel ↓ δ_1 ↑		AbsRel↓ δ_1 ↑	
DiverseDepth [18] ResNeXt50 [104]	25M	1.3M	320K	0.117	0.875	0.190	0.704	0.376	0.631	0.108	0.882	0.228	0.694
MiDaS [27]ResNeXt101	88M	1.3M	2M	0.111	0.885	0.236	0.630	0.332	0.715	0.111	0.886	0.184	0.752
Leres [25]ResNeXt101		1.3M	354K	0.090	0.916	0.149	0.784	0.271	0.766	0.095	0.912	0.171	0.777
Omnidata [35]ViT-Base		1.3M	12.2M	0.074	0.945	0.149	0.835	0.339	0.742	0.077	0.935	0.166	0.778
HDN [29]ViT-Large [86]	306M	1.3M	300K	0.069	0.948	0.115	0.867	0.246	0.780	0.080	0.939	0.121	0.833
DPT-large [28]ViT-Large		1.3M	188K	0.098	0.903	0.100	0.901	0.182	0.758	0.078	0.938	0.078	0.946
DepthAnything [28]ViT-Large		142M	63.5M	0.043	0.981	0.076	0.947	0.277	0.759	0.042	0.980	0.127	0.882
Marigold [28]Latent diffusion V2 [87]	899M	5B	74K	0.055	0.961	0.099	0.916	0.308	0.773	0.064	0.951	0.065	0.960
Ours CSTM labelViT-Small	22M	142M	16M	0.056	0.965	0.064	0.950	0.247	0.789	0.033†	0.985	0.062	0.955
Ours CSTM_imageConvNeXt-Large [102]	198M	14.2M	8M	0.058	0.963	0.053	0.965	0.211	0.825	0.074	0.942	0.064	0.965
Ours CSTM_labelConvNeXt-Large		14.2M	8M	0.050	0.966	0.058	0.970	0.224	0.805	0.074	0.941	0.066	0.964
Ours CSTM labelViT-Large	306M	142M	16M	0.042	0.980	0.046	0.979	0.141	0.882	0.021†	0.993	0.042	0.987
Ours CSTM_labelViT-giant [105]	1011M	142M	16M	0.043	0.981	0.044	0.982	0.136	0.895	0.022†	0.994	0.042	0.983

方法主干网络	# 参数	# 数据		纽约大学深度数据集 v2(NYUv2)		基带视觉基准数据集 (KITTI)		二极管深度数据集 (完整)(DIODE(Full))		扫描网络数据集 (ScanNet)		苏黎世联邦理工学院 3D 数据集 (ETH3D)	
		预训练	训练	绝对相对误差 ↓ δ_1 ↑		绝对相对误差 ↓ δ_1 ↑		AbsRel ↓ δ_1 ↑		AbsRel ↓ δ_1 ↑		绝对相对误差 ↓ δ_1 ↑	
多样深度模型 [18] ResNeXt50 [104]	25M	1.3M	320K	0.117	0.875	0.190	0.704	0.376	0.631	0.108	0.882	0.228	0.694
多尺度密集注意力网络 [27] 残差网络增强版 101(MiDaS [27]ResNeXt101)	88M	1.3M	2M	0.111	0.885	0.236	0.630	0.332	0.715	0.111	0.886	0.184	0.752
分层边缘细化网络 [25] 残差网络增强版 101(Leres [25]ResNeXt101)		1.3M	354K	0.090	0.916	0.149	0.784	0.271	0.766	0.095	0.912	0.171	0.777
全数据模型 [35] 视觉变换器基础版 (Omnidata [35]ViT - Base)		1.3M	12.2M	0.074	0.945	0.149	0.835	0.339	0.742	0.077	0.935	0.166	0.778
混合深度网络 [29] 视觉变换器大模型 [86](HDN [29]ViT - Large [86])	306M	1.3M	300K	0.069	0.948	0.115	0.867	0.246	0.780	0.080	0.939	0.121	0.833
密集预测变换器大模型 [28] 视觉变换器大模型 (DPT - large [28]ViT - Large)		1.3M	188K	0.098	0.903	0.100	0.901	0.182	0.758	0.078	0.938	0.078	0.946
任意深度模型 [28] 视觉变换器大模型 (DepthAnything [28]ViT - Large)		142M	63.5M	0.043	0.981	0.076	0.947	0.277	0.759	0.042	0.980	0.127	0.882
金盏花模型 [28] 潜在扩散模型 V2 [87](Marigold [28][Latent diffusion V2 [87])	899M	5B	74K	0.055	0.961	0.099	0.916	0.308	0.773	0.064	0.951	0.065	0.960
我们的基于条件语义转换模块的标签视觉变换器小模型 (Ours CSTM labelViT - Small)	22M	142M	16M	0.056	0.965	0.064	0.950	0.247	0.789	0.033†	0.985	0.062	0.955
我们的基于条件语义转换模块的图像卷积神经网络大模型 [102](Ours CSTM_imageConvNeXt - Large [102])	198M	14.2M	8M	0.058	0.963	0.053	0.965	0.211	0.825	0.074	0.942	0.064	0.965
我们的基于条件语义转换模块的标签卷积神经网络大模型 (Ours CSTM_labelConvNeXt - Large)		14.2M	8M	0.050	0.966	0.058	0.970	0.224	0.805	0.074	0.941	0.066	0.964
我们的基于条件语义转换模块的标签视觉变换器大模型 (Ours CSTM labelViT - Large)	306M	142M	16M	0.042	0.980	0.046	0.979	0.141	0.882	0.021†	0.993	0.042	0.987
我们的基于条件语义转换模块的标签视觉变换器巨型模型 [105](Ours CSTM_labelViT - giant [105])	1011M	142M	16M	0.043	0.981	0.044	0.982	0.136	0.895	0.022†	0.994	0.042	0.983

† : ScanNet is partly annotated with normal [40]. For samples without normal annotations, these models use depth labels to facilitate normal learning.

† :ScanNet(扫描网络数据集) 部分标注了法线信息 [40]。对于没有法线标注的样本，这些模型使用深度标签来促进法线学习。

Furthermore, We collect 5 unseen datasets to do more metric accuracy evaluation. These datasets contain a wide range of indoor and outdoor scenes, including rooms, buildings, and driving scenes. The camera models are also varied. We mainly compare with the SoTA metric depth estimation methods and take their NYUv2 and KITTI models for indoor and outdoor scene evaluation respectively. From Tab. 3, we observe that although NuScenes is similar to KITTI, existing methods face a noticeable performance decrease. In contrast, our model is more robust.

此外，我们收集了 5 个未见数据集以进行更准确的指标评估。这些数据集包含广泛的室内和室外场景，包括房间、建筑物和驾驶场景。相机型号也多种多样。我们主要与最先进的度量深度估计方法进行比较，并分别采用它们的 NYUv2(纽约大学数据集版本 2) 和 KITTI(卡尔斯鲁厄理工学院和丰田工业大学联合数据集) 模型进行室内和室外场景评估。从表 3 中我们可以观察到，尽管 NuScenes(努场景数据集) 与 KITTI 相似，但现有方法的性能出现了明显下降。相比之下，我们的模型更具鲁棒性。

Generalization over diverse scenes. Affine-invariant depth benchmarks decouple the scale's effect, which aims to evaluate the model's generalization ability to diverse scenes. Recent impact works, such as MiDaS, LeReS, DPT, Marigold, and DepthAnything achieved promising performance on them. Following them, we test on 5 datasets and manually align the scale and shift to the ground-truth depth before evaluation. Results are reported in Tab. 4. Although our method enforces the network to recover the more challenging metric, our method outperforms them on all datasets.

不同场景的泛化能力。仿射不变深度基准消除了尺度的影响，旨在评估模型对不同场景的泛化能力。最近有影响力的工作，如 MiDaS(多尺度密集预测)、LeReS(轻量级残差网络)、DPT(密集预测变换器)、Marigold(金盏花) 和 DepthAnything(任意深度) 在这些基准上取得了不错的性能。效仿它们，我们在 5 个数据集上进行测试，并在评估前手动将尺度和偏移与真实深度对齐。结果报告在表 4 中。尽管我们的方法要求网络恢复更具挑战性的度量，但我们的方法在所有数据集上都优于它们。

Evaluation on surface normal benchmarks. We evaluate our methods on ScanNet, NYU, and iBims-1 surface normal benchmarks. Results are reported in Tab. 2. Firstly, we organize a zero-shot testing benchmark on NYU dataset, see methods denoted with 'ZS' in the table. We compare with existing methods which are trained on ScanNet or Taskonomy and have achieved promising performance on them, such as Polymax [34] and Bae et al. [33]. Our method surpasses them over most metrics. Cmparing with methods that have been overfitted the NYU data domain for hundreds of epochs (marked with blue), our zero-shot testing outperforms them on all metrics. Our post-finetuned models ('FT' marks) further boost the performance. Similarly, we also achieve SoTA performance on iBims-1 and Scannet benchmarks. For the iBims-1 dataset, we follow IronDepth [42] to generate the ground-truth normal annotations.

表面法线基准评估。我们在 ScanNet(扫描网络数据集)、NYU(纽约大学数据集) 和 iBims - 1(室内建筑材料表面法线数据集 1) 表面法线基准上评估我们的方法。结果报告在表 2 中。首先，我们在 NYU 数据集上组织了一个零样本测试基准，见表中标记为“ZS”的方法。我们与在 ScanNet 或 Taskonomy(任务 onomy 数据集) 上训练并在这些数据集上取得良好性能的现有方法进行比较，如 Polymax [34] 和 Bae 等人 [33]。我们的方法在大多数指标上超过了它们。与在 NYU 数据域上过度拟合数百个训练周期的方法 (标记为蓝色) 相比，我们的零样本测试在所有指标上都优于它们。我们的后微调模型 (标记为“FT”) 进一步提升了性能。同样，我们在 iBims - 1 和 Scannet 基准上也取得了最先进的性能。对于 iBims - 1 数据集，我们遵循 IronDepth [42] 的方法生成真实法线标注。

TABLE 5 - Training and testing datasets used for experiments.

表 5 - 实验使用的训练和测试数据集。

Datasets	Scenes	Source	Label	Size	#Cam.
Training Data					
DDAD [114]	Outdoor	Real-world	Depth	80K	36+
Lyft [115]	Outdoor	Real-world	Depth	$\sim$ 50K	6+
Driving Stereo (DS) [116]	Outdoor	Real-world	Depth	$\sim$ 181 K	1
DIML [117]	Outdoor	Real-world	Depth	122K	10
Arogoverse2 [118]	Outdoor	Real-world	Depth	3515K	6+
Cityscapes [119]	Outdoor	Real-world	Depth	170K	1
DSEC [120]	Outdoor	Real-world	Depth	26K	1
Mapillary PSD [89]	Outdoor	Real-world	Depth	750K	1000+
Pandaset [121]	Outdoor	Real-world	Depth	48K	6
UASOL [122]	Outdoor	Real-world	Depth	1370K	1
Virtual KITTI [123]	Outdoor	Synthesized	Depth	37K	2
Waymo [124]	Outdoor	Real-world	Depth	1M	5
Matterport3d [125]	In/Out	Real-world	Depth + Normal	144K	3
Taskonomy [125]	Indoor	Real-world	Depth + Normal	4M	$\sim$ 1M
Replica [126]	Indoor	Real-world	Depth + Normal	150K	1
ScanNet † [107]	Indoor	Real-world	Depth + Normal	2.5M	1
HM3d [127]	Indoor	Real-world	Depth + Normal	2000K	1
Hypersim [128]	Indoor	Synthesized	Depth + Normal	54K	1
Testing Data					
NYU [58]	Indoor	Real-world	Depth+Normal	654	1
KITTI [59]	Outdoor	Real-world	Depth	652	4
ScanNet † [107]	Indoor	Real-world	Depth+Normal	700	1
NuScenes (NS) [108]	Outdoor	Real-world	Depth	10K	6
ETH3D [129]	Outdoor	Real-world	Depth	431	1
DIODE [110]	In/Out	Real-world	Depth	771	1
iBims-1 [109]	Indoor	Real-world	Depth	100	1

数据集	场景	来源	标签	大小	摄像头数量
训练数据					
DDAD [114]	室外	真实世界	深度	80K	36+
Lyft [115]	室外	真实世界	深度	∼ 50K	6+
驾驶立体数据集 (Driving Stereo，DS) [116]	室外	真实世界	深度	∼ 181 K	1
DIML [117]	室外	真实世界	深度	122K	10
Arogoverse2 [118]	室外	真实世界	深度	3515K	6+
城市景观数据集 (Cityscapes) [119]	室外	真实世界	深度	170K	1
DSEC [120]	室外	真实世界	深度	26K	1
Mapillary PSD [89]	室外	真实世界	深度	750K	1000+
Pandaset [121]	室外	真实世界	深度	48K	6
UASOL [122]	室外	真实世界	深度	1370K	1
虚拟基蒂数据集 (Virtual KITTI) [123]	室外	合成的	深度	37K	2
Waymo [124]	室外	真实世界	深度	1M	5
Matterport3d [125]	室内/室外	真实世界	深度 + 法线	144K	3
任务分类数据集 (Taskonomy) [125]	室内	真实世界	深度 + 法线	4M	∼ 1M
Replica [126]	室内	真实世界	深度 + 法线	150K	1
扫描网络数据集 (ScanNet) † [107]	室内	真实世界	深度 + 法线	2.5M	1
HM3d [127]	室内	真实世界	深度 + 法线	2000K	1
Hypersim [128]	室内	合成的	深度 + 法线	54K	1
测试数据					
纽约大学数据集 (NYU) [58]	室内	真实世界	深度 + 法线	654	1
基蒂数据集 (KITTI) [59]	室外	真实世界	深度	652	4
扫描网络数据集 (ScanNet) † [107]	室内	真实世界	深度 + 法线	700	1
NuScenes(NS) [108]	室外	真实世界	深度	10K	6
ETH3D [129]	室外	真实世界	深度	431	1
DIODE [110]	室内/室外	真实世界	深度	771	1
iBims - 1 [109]	室内	真实世界	深度	100	1

† ScanNet is a non-zero-shot testing dataset for our ViT models.

† ScanNet(扫描网络) 是我们的视觉 Transformer(ViT) 模型的非零样本测试数据集。

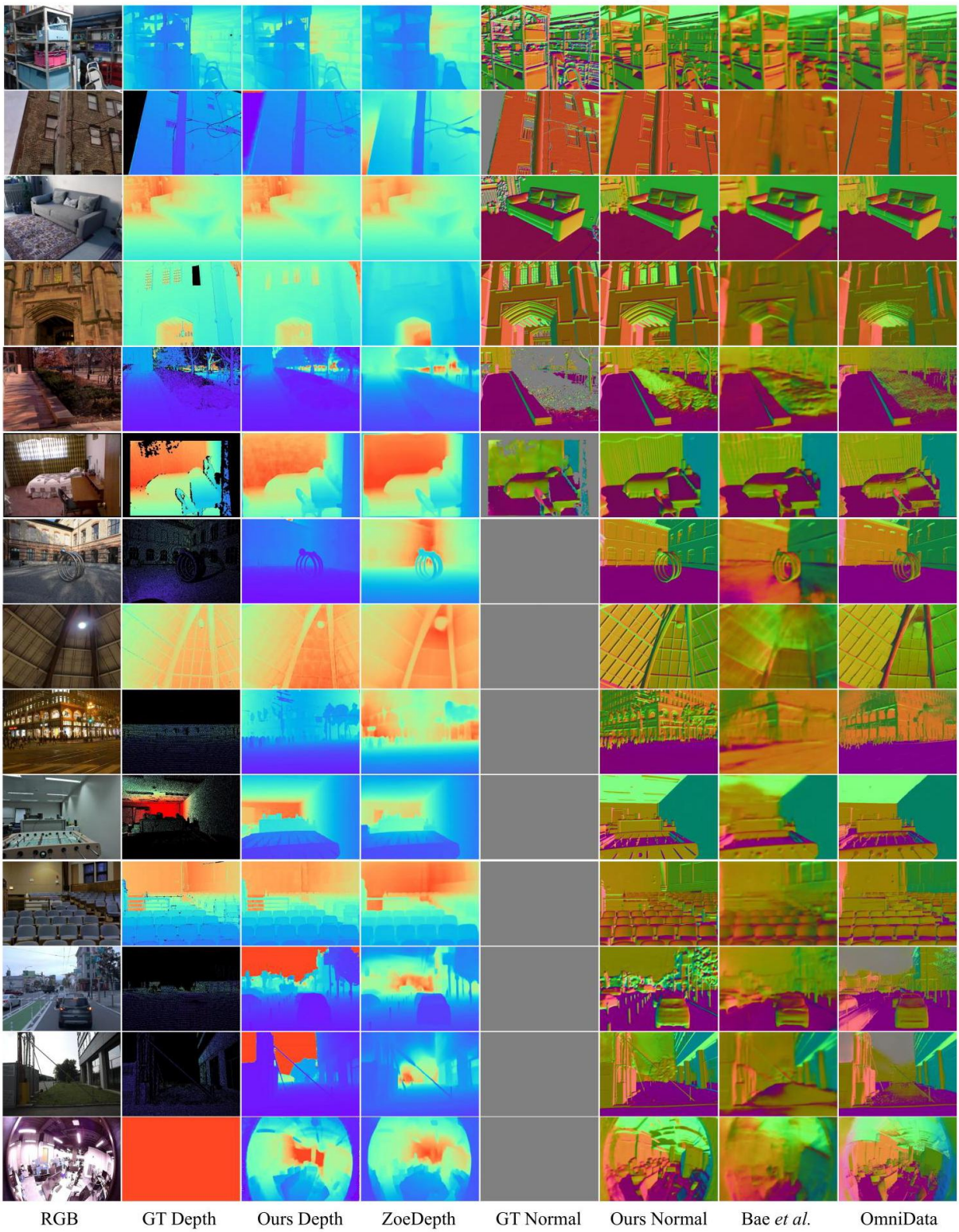

Fig. 11 - Qualitative comparisons of metric depth and surface normals for iBims, DIODE, NYU, Eth3d, Nuscenes, and self-collected drone datasets. We present visualization results of our predictions ('Ours Depth' / 'Ours Normal'), groundtruth labels ('GT Depth' / 'GT Normal') and results from other metric depth ('ZoeDepth' [36]) and surface normal methods ('Bae et al.' [33] and 'OmniData' [35]).

图 11 - 针对 iBims、DIODE、NYU、Eth3d、Nuscenes 和自行采集的无人机数据集的度量深度和表面法线的定性比较。我们展示了我们的预测结果 (“我们的深度”/“我们的法线”)、真实标签 (“真实深度”/“真实法线”) 以及其他度量深度方法 (“ZoeDepth [36]”) 和表面法线方法 (“Bae 等人 [33]”和“OmniData [35]”) 的可视化结果。

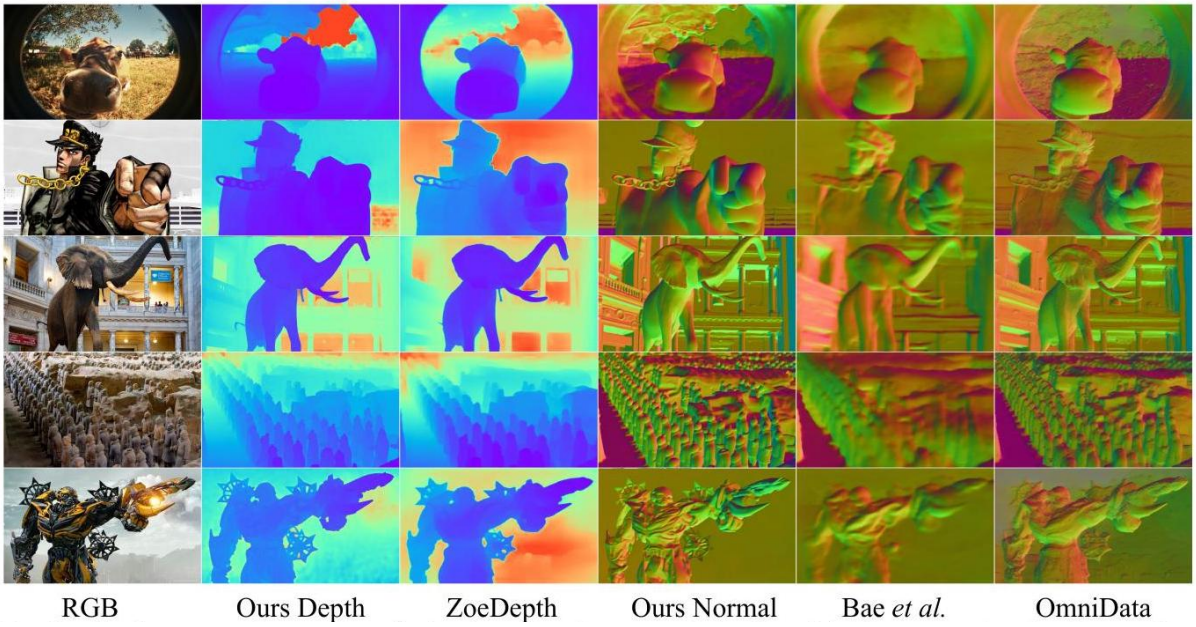

Fig. 12 - Qualitative comparisons of metric depth and surface normals in the wild. We present visualization results of our predictions ('Ours Depth' / 'Ours Normal') and results from other metric depth ('ZoeDepth' [36]) and surface normal methods ('Bae et al.' [33] and 'OmniData' [35]).

图 12 - 野外场景下度量深度和表面法线的定性比较。我们展示了我们的预测结果 (“我们的深度”/“我们的法线”) 以及其他度量深度方法 (“ZoeDepth [36]”) 和表面法线方法 (“Bae 等人 [33]”和“OmniData [35]”) 的可视化结果。

4.1 Zero-shot Generalization

4.1 零样本泛化

Qualitative comparisons of surface normals and depths.

表面法线和深度的定性比较。

We visualize our predictions in Fig. 11. A comparison with another widely used generalized metric depth method, ZoeDepth [36], demonstrates that our approach produces depth maps with superior details on fine-grained structures (objects in row1, suspension lamp in row4, beam in row8), and better foreground/background distinction (row 11, 12). In terms of surface normal prediction, our normal maps exhibits significantly finer details compared to Bae. et al. [33] and can handle some cases where their method fail (row7, 8, 9) . Our method not only generalizes well across diverse scenarios but can also be directly applied to unseen camera models like the fisheye camera

shown in row 12. More visualization results for in-the-wild images are presented in Fig. 12, including comic-style (Row 2) and CG(computer graphics)-generated objects (Row5)

我们在图 11 中可视化了我们的预测结果。与另一种广泛使用的广义度量深度方法 ZoeDepth [36] 进行比较，结果表明我们的方法在细粒度结构 (第一行的物体、第四行的吊灯、第八行的横梁) 上生成的深度图具有更优的细节，并且前景/背景区分更好 (第 11、12 行)。在表面法线预测方面，与 Bae 等人 [33] 的方法相比，我们的法线图显示出明显更精细的细节，并且能够处理他们的方法失效的一些情况 (第 7 行，$8, 9$))。我们的方法不仅在不同场景下具有良好的泛化能力，还可以直接应用于未见过的相机模型，如第 12 行所示的鱼眼相机。图 12 展示了更多野外图像的可视化结果，包括漫画风格 (第 2 行) 和计算机图形 (CG) 生成的物体 (第 5 行)。

4.2 Applications Based on Our Method

4.2 基于我们方法的应用

We apply the CSTM_image model to various tasks.

我们将 CSTM_image 模型应用于各种任务。

3D scene reconstruction . To present our method's ability to recover real-world metric 3D , we first conduct a quantitative comparison on 9 unseen NYUv2 scenes. We predict per-frame metric depth and fuse these with the provided camera poses, with results detailed in Table 6. We compare our approach to several methods: the video consistent depth prediction method (RCVD [130]), unsupervised video depth estimation (SC-DepthV2 [131]), 3D scene shape recovery (LeReS [25]), affine-invariant depth estimation (DPT [28]), and multi-view stereo reconstruction (DPSNet [48], Sim-pleRecon [132]). Except for the multi-view approaches and our method, all others require aligning scales with ground truth depth for each frame. While our approach is not specifically designed for video or multi-view reconstruction, it demonstrates promising frame consistency and significantly more accurate 3D scene reconstructions in these zero-shot scenarios. Qualitative comparisons in Fig. 13 reveal that our reconstructions exhibit considerably less noise and fewer outliers.

3D 场景重建。为了展示我们的方法恢复真实世界度量 3D 的能力，我们首先对 9 个未见过的 NYUv2 场景进行了定量比较。我们预测每帧的度量深度，并将其与提供的相机位姿进行融合，结果详见表 6。我们将我们的方法与几种方法进行了比较: 视频一致深度预测方法 (RCVD [130])、无监督视频深度估计 (SC - DepthV2 [131])、3D 场景形状恢复 (LeReS [25])、仿射不变深度估计 (DPT [28]) 和多视图立体重建 (DPSNet [48]、SimpleRecon [132])。除了多视图方法和我们的方法外，其他所有方法都需要为每一帧与真实深度进行尺度对齐。虽然我们的方法并非专门为视频或多视图重建而设计，但在这些零样本场景中，它展示了有前景的帧一致性和显著更准确的 3D 场景重建效果。图 13 中的定性比较表明，我们的重建结果噪声明显更少，异常值也更少。

Dense-SLAM mapping. Monocular SLAM is a key robotic application that uses a single video input to create trajectories and dense 3D maps. However, due to limited photometric and geometric constraints, existing methods struggle with scale drift in large scenes and fail to recover accurate metric information. Our robust metric depth estimation serves as a strong depth prior for the SLAM system. To demonstrate this, we input our metric depth into

the state-of-the-art SLAM system, Droid-SLAM [37], and evaluate the trajectory on KITTI without any tuning. Results are shown in Table 7. With access to accurate per-frame metric depth, Droid-SLAM experiences a significant reduction in translation drift (t_{rel}). Additionally, our depth data enables Droid-SLAM to achieve denser and more precise 3D mapping, as illustrated in Fig. 3 and detailed in the supplementary materials.

稠密同步定位与地图构建 (Dense - SLAM) 映射。单目 SLAM 是一种关键的机器人应用，它使用单个视频输入来创建轨迹和稠密 3D 地图。然而，由于有限的光度和几何约束，现有方法在大场景中难以处理尺度漂移问题，并且无法恢复准确的度量信息。我们强大的度量深度估计为 SLAM 系统提供了强大的深度先验。为了证明这一点，我们将我们的度量深度输入到最先进的 SLAM 系统 Droid - SLAM [37] 中，并在 KITTI 数据集上进行轨迹评估，且未进行任何调优。结果如表 7 所示。由于能够获取准确的每帧度量深度，Droid - SLAM 的平移漂移 (t_{rel}) 显著减少。此外，我们的深度数据使 Droid - SLAM 能够实现更密集、更精确的 3D 映射，如图 3 所示，并在补充材料中详细说明。

We also tested on the ETH3D SLAM benchmarks, with results in Table 8. Using our metric depth predictions, Droid-SLAM shows improved performance, although the gains are less pronounced in the smaller indoor scenes of ETH3D compared to KITTI.

我们还在 ETH3D SLAM 基准测试上进行了测试，结果如表 8 所示。使用我们的度量深度预测结果，Droid - SLAM 的性能有所提升，不过与 KITTI 数据集相比，在 ETH3D 的较小室内场景中，提升效果不太明显。

Metrology in the wild. To demonstrate the robustness and accuracy of our recovered metric 3D shapes, we downloaded Flickr photos taken by various cameras and extracted coarse camera intrinsic parameters from their meta-data. We utilized our CSTM_image model to reconstruct the metric shapes and measure the sizes of structures (marked in red in Fig. 14), with ground-truth sizes shown in blue. The results indicate that our measured sizes closely align with the ground-truth values.

野外计量。为了证明我们恢复的度量 3D 形状的鲁棒性和准确性，我们下载了各种相机拍摄的 Flickr 照片，并从其元数据中提取了粗略的相机内参。我们利用我们的 CSTM_image 模型重建度量形状，并测量结构的尺寸 (在图 14 中用红色标记)，真实尺寸用蓝色显示。结果表明，我们测量的尺寸与真实值非常接近。

Monocular reconstruction in the wild. To further visualize the reconstruction quality of our recovered metric depth, we randomly collect images from the internet and recover their metric 3D and normals. As there is no focal length provided, we select proper focal lengths according to the reconstructed shape and normal maps. The reconstructed pointclouds are colorized by their corresponding normals (Different views are marked by red and orange arrays in Fig. 15).

野外单目重建。为了进一步可视化我们恢复的度量深度的重建质量，我们从互联网上随机收集图像，并恢复它们的度量 3D 和法线。由于没有提供焦距，我们根据重建的形状和法线图选择合适的焦距。重建的点云通过其相应的法线进行着色 (不同视角在图 15 中用红色和橙色箭头标记)。

TABLE 6 - Quantitative comparison of 3D scene reconstruction with LeReS [25], DPT [28], RCVD [130], SC-DepthV2 [131], and two learning-based MVS methods (DPSNet [48], SimpleRecon [132]) on 9 unseen NYUv2 scenes. Apart from the MVS approaches and ours, other methods have to align the scale with ground truth depth for each frame. As a result, our reconstructed 3D scenes achieve the best performance.

表 6 - 在 9 个未见的 NYUv2 场景上，将 3D 场景重建与 LeReS [25]、DPT [28]、RCVD [130]、SC - DepthV2 [131] 以及两种基于学习的多视图立体 (MVS) 方法 (DPSNet [48]、SimpleRecon [132]) 进行定量比较。除了多视图立体方法和我们的方法外，其他方法必须为每一帧与真实深度进行尺度对齐。因此，我们重建的 3D 场景取得了最佳性能。

Method		Basement_0001a		Bedroom_0015	Dining_roo C- l_1 ↓	F-score ↑	Kitchen_0008 C - l_1 ↓ F F-score ↑		Classroom_0004Playroom_0002				Office_0024		Office_0004		Dining_room_	
	C - l_1 ↓	F-score ↑	C- l_1 ↓	7-score ↑					C- l_1 ↓	F-score ↑	C- l_1 ↓	F-score ↑	C- l_1 ↓	7-score ↑	C- l_1 ↓ F	F-score ↑	C- l_1 ↓	F-score ↑
RCVD [130]	0.364	0.276	0.074	0.582	0.462	0.251	0.053	0.620	0.187	0.327	0.791	0.187	0.324	0.241	0.646	0.217	0.445	0.253
SC-DepthV2 [131]	0.254	0.275	0.064	0.547	0.749	0.229	0.049	0.624	0.167	0.267	0.426	0.263	0.482	0.138	0.516	0.244	0.356	0.247
DPSNet [48]	0.243	0.299	0.195	0.276	0.995	0.186	0.269	0.203	0.296	0.195	0.141	0.485	0.199	0.362	0.210	0.462	0.222	0.493
DPT [25]	0.698	0.251	0.289	0.226	0.396	0.364	0.126	0.388	0.780	0.193	0.605	0.269	0.454	0.245	0.364	0.279	0.751	0.185
LeReS [25]	0.081	0.555	0.064	0.616	0.278	0.427	0.147	0.289	0.143	0.480	0.145	0.503	0.408	0.176	0.096	0.497	0.241	0.325
SimpleRecon [132]	0.068	0.695	0.086	0.449	0.199	0.413	0.055	0.624	0.142	0.461	0.092	0.517	0.054	0.638	0.051	0.681	0.165	0.565
Ours	0.042	0.736	0.059	0.610	0.159	0.485	0.050	0.645	0.145	0.445	0.036	0.814	0.069	0.638	0.045	0.700	0.060	0.663

方法		地下室 _0001a		卧室 _0015	餐厅 C- l_1 ↓	F 值↑	厨房 _0008 C - l_1 ↓ F F 值↑		教室 _0004 游戏室 _0002				办公室 _0024		办公室 _0004		餐厅 _	
	C - l_1 ↓	F 值↑	C- l_1 ↓	7 值↑					C- l_1 ↓	F 值↑	C- l_1 ↓	F 值↑	C- l_1 ↓	7 值↑	C- l_1 ↓ F	F 值↑	C- l_1 ↓	F 值↑
RCVD [130]	0.364	0.276	0.074	0.582	0.462	0.251	0.053	0.620	0.187	0.327	0.791	0.187	0.324	0.241	0.646	0.217	0.445	0.253
SC 深度 V2 [131]	0.254	0.275	0.064	0.547	0.749	0.229	0.049	0.624	0.167	0.267	0.426	0.263	0.482	0.138	0.516	0.244	0.356	0.247
DPS 网络 [48]	0.243	0.299	0.195	0.276	0.995	0.186	0.269	0.203	0.296	0.195	0.141	0.485	0.199	0.362	0.210	0.462	0.222	0.493
DPT [25]	0.698	0.251	0.289	0.226	0.396	0.364	0.126	0.388	0.780	0.193	0.605	0.269	0.454	0.245	0.364	0.279	0.751	0.185
LeReS [25]	0.081	0.555	0.064	0.616	0.278	0.427	0.147	0.289	0.143	0.480	0.145	0.503	0.408	0.176	0.096	0.497	0.241	0.325
简单重建 [132]	0.068	0.695	0.086	0.449	0.199	0.413	0.055	0.624	0.142	0.461	0.092	0.517	0.054	0.638	0.051	0.681	0.165	0.565
我们的方法	0.042	0.736	0.059	0.610	0.159	0.485	0.050	0.645	0.145	0.445	0.036	0.814	0.069	0.638	0.045	0.700	0.060	0.663

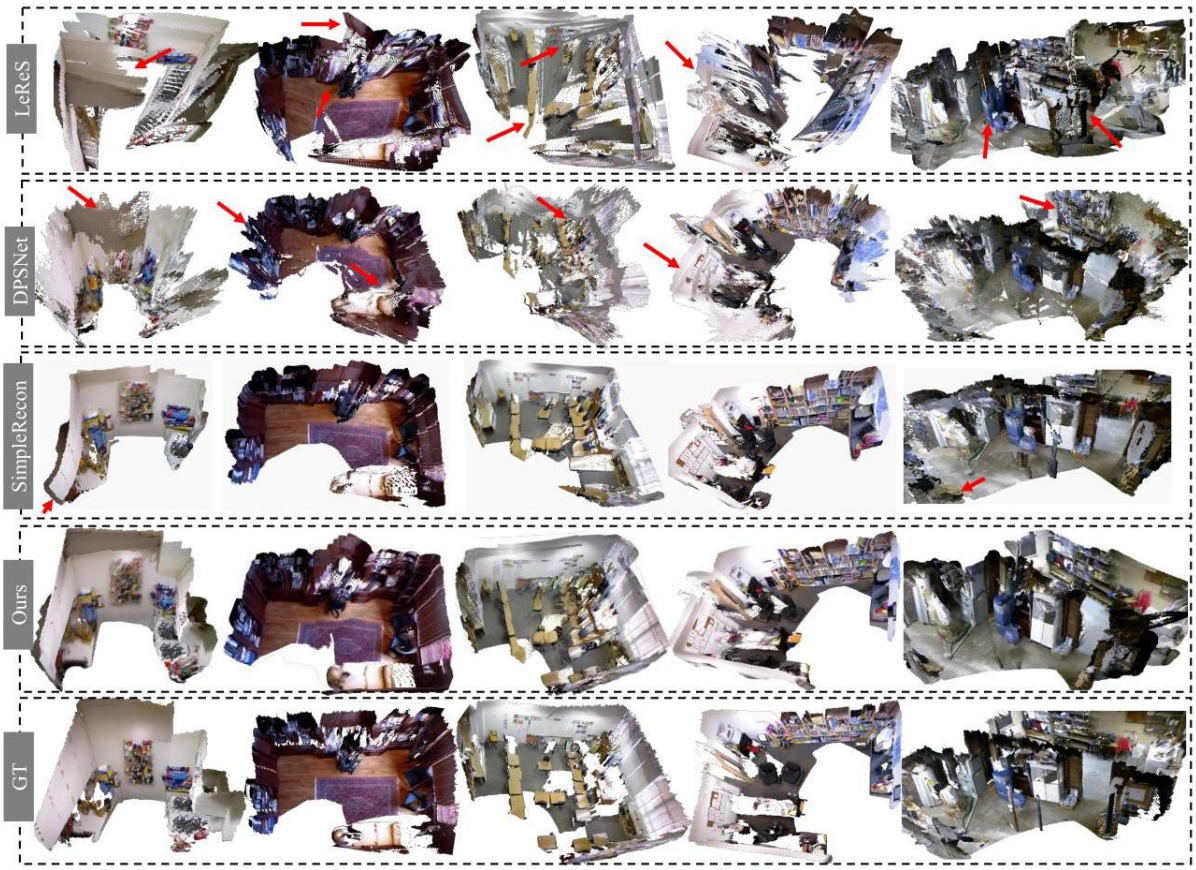

Fig. 13 - Reconstruction of zero-shot scenes with multiple views. We sample several NYUv2 scenes for 3D reconstruction comparison. As our method can predict accurate metric depth, thus all frame's predictions are fused directly for reconstruction. By contrast, LeReS [25]'s depth is up to an unknown scale and shift, causing noticeable distortions. For MVS methods, DPSNet [48] fails on low-texture backgrounds, while SimpleRecon [132] distorts regions without sufficient observations.

图 13 - 多视图零样本场景重建。我们选取了几个 NYUv2 场景进行三维重建比较。由于我们的方法可以预测出准确的度量深度，因此所有帧的预测结果可直接融合进行重建。相比之下，LeReS [25] 的深度存在未知的尺度和偏移，会导致明显的失真。对于多视图立体 (MVS) 方法，DPSNet [48] 在低纹理背景下会失效，而 SimpleRecon [132] 在观测不足的区域会出现失真。

TABLE 7 - Comparison with SoTA SLAM methods on KITTI. We input predicted metric depth to the Droid-SLAM [37] ('Droid+Ours'), which outperforms others by a large margin on trajectory accuracy.

表 7 - 在 KITTI 数据集上与最先进 (SoTA) 的同时定位与地图构建 (SLAM) 方法的比较。我们将预测的度量深度输入到 Droid - SLAM [37]('Droid+Ours') 中，该方法在轨迹精度上大幅优于其他方法。

Method	Seq 00	Seq 02	Seq 05	Seq 06	Seq 08	Seq 09	Seq 10
	Translational RMS drift (t_{rel}, ↓) / Rotational RMS drift (r_{rel}, ↓)						
GeoNet [133]	27.6 / 5.72	42.24 / 6.14	20.12/7.67	9.28/4.34	18.59 / 7.85	23.94 / 9.81	20.73/9.1
VISO2-M [134]	12.66/2.73	9.47 / 1.19	15.1 / 3.65	6.8/1.93	14.82 / 2.52	3.69 / 1.25	21.01 /3.26
ORB-V2 [12]	11.43/0.58	10.34/0.26	9.04/0.26	14.56/0.26	11.46/0.28	9.3/0.26	2.57 / 0.32
Droid [37]	33.9/0.29	34.88 / 0.27	23.4 / 0.27	17.2/0.26	39.6/0.31	21.7 / 0.23	7/0.25
Droid+Ours	1.44/0.37	2.64/0.29	**1.44**/**0.25**	0.6/0.2	2.2/0.3	**1.63**/**0.22**	2.73/0.23

方法	序列 00	序列 02	序列 05	序列 06	序列 08	序列 09	序列 10
	平移均方根漂移 (t_{rel}, ↓) / 旋转均方根漂移 (r_{rel}, ↓)						
地理网络 (GeoNet) [133]	27.6 / 5.72	42.24 / 6.14	20.12/7.67	9.28/4.34	18.59 / 7.85	23.94 / 9.81	20.73/9.1
VISO2 - M [134]	12.66/2.73	9.47 / 1.19	15.1 / 3.65	6.8/1.93	14.82 / 2.52	3.69 / 1.25	21.01 /3.26
ORB - V2 [12]	11.43/0.58	10.34/0.26	9.04/0.26	14.56/0.26	11.46/0.28	9.3/0.26	2.57 / 0.32
机器人 (Droid) [37]	33.9/0.29	34.88 / 0.27	23.4 / 0.27	17.2/0.26	39.6/0.31	21.7 / 0.23	7/0.25
机器人 (Droid)+ 我们的方法	1.44/0.37	2.64/0.29	**1.44**/**0.25**	0.6/0.2	2.2/0.3	**1.63**/**0.22**	2.73/0.23

TABLE 8 - Comparison of VO error on ETH3D benchmark. Droid SLAM system is input with our depth ('Droid + Ours'), and ground-truth depth (' Droid + GT'). The average trajectory error is reported.

表 8 - ETH3D 基准上视觉里程计 (VO) 误差的比较。将我们的深度数据 ("Droid + 我们的数据") 和真实深度数据 ("Droid + 真实值") 分别输入到 Droid SLAM 系统中。报告了平均轨迹误差。

	Einstein global	Manquin4	Motion1	Plant- scene3	sfm_house _loop	sfm_lab _room2
	Average trajectory error (↓)					
Droid	4.7	0.88	0.83	0.78	5.64	0.55
Droid + Ours	1.5	0.69	0.62	0.34	4.03	0.53
Droid + GT	0.7	0.006	0.024	0.006	0.96	0.013

	爱因斯坦全局 (Einstein global)	曼奎因 4(Manquin4)	运动 1(Motion1)	植物 - 场景 3(Plant - scene3)	结构从运动房屋循环 (sfm_house _loop)	结构从运动实验室房间 2(sfm_lab _room2)
	平均轨迹误差 (↓) (Average trajectory error (↓))					
机器人 (Droid)	4.7	0.88	0.83	0.78	5.64	0.55
机器人 + 我们的方法 (Droid + Ours)	1.5	0.69	0.62	0.34	4.03	0.53
机器人 + 真实值 (Droid + GT)	0.7	0.006	0.024	0.006	0.96	0.013

4.3 Ablation Study

4.3 消融研究

Ablation on canonical transformation. We examine the impact of our proposed canonical transformations for input images (CSTM_input) and ground-truth labels (CSTM_output). Results are presented in Table 9. We trained the model on a mixed dataset of 90,000 images and tested it across six datasets. A naive baseline (Ours w/o CSTM) removes the CSTM modules, enforcing the same supervision as our approach. Without CSTM, the model struggles to converge on mixed metric datasets and fails to achieve metric predictions on zero-shot datasets. This limitation is why recent mixed-data training methods often resort to learning affine-invariant depth to sidestep metric challenges. In

正则变换的消融实验。我们研究了所提出的针对输入图像的正则变换 (CSTM_input) 和真实标签的正则变换 (CSTM_output) 的影响。结果见表 9。我们在包含 90,000 张图像的混合数据集上训练模型，并在六个数据集上进行测试。一个简单的基线模型 (Ours w/o CSTM) 移除了 CSTM 模块，采用与我们方法相同的监督方式。没有 CSTM 时，模型在混合度量数据集上难以收敛，并且无法在零样本数据集上实现度量预测。这一局限性正是近期混合数据训练方法常常采用学习仿射不变深度来规避度量挑战的原因。在

Fig. 14 - Metrology of in-the-wild scenes. We collect several Flickr photos, which are captured by various cameras. With photos' metadata, we reconstruct the 3D metric shape and measure structures' sizes. Red and blue marks are ours and ground-truth sizes respectively.

图 14 - 自然场景的度量。我们收集了几张 Flickr 照片，这些照片由不同的相机拍摄。利用照片的元数据，我们重建了三维度量形状并测量了物体结构的尺寸。红色和蓝色标记分别表示我们的测量结果和真实尺寸。

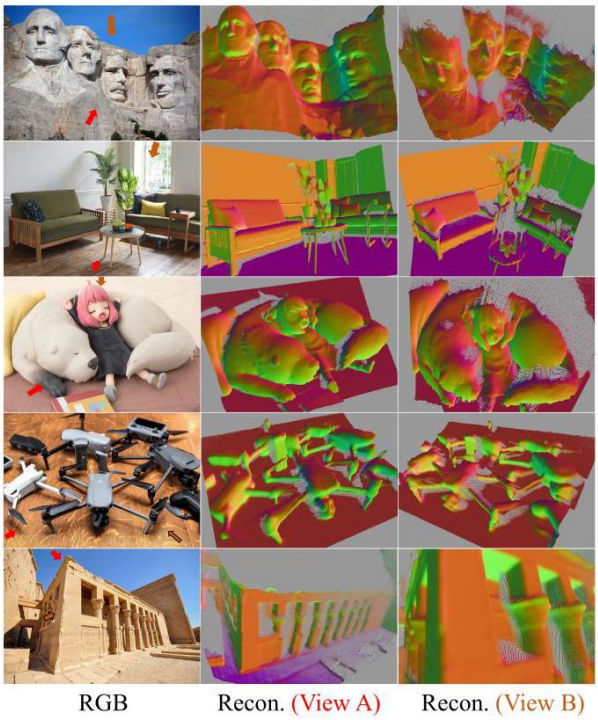

Fig. 15 - Reconstruction from in-the-wild single images. We collect web images and select proper focal lengths. The reconstructed pointclouds are colorized by normals.

图 15 - 从自然单张图像进行重建。我们收集网络图像并选择合适的焦距。重建的点云通过法线进行着色。

contrast, both of our CSTM methods enable the model to attain metric prediction capabilities and achieve comparable performance. Table 1 confirms this comparable performance. Thus, adjusting supervision and the appearance of input images during training effectively addresses metric ambiguity issues. Additionally, we compared our approach with CamConvs [38], which incorporates the camera model in the decoder using a 4-channel feature. While CamConvs uses the same training schedule, model, and data, it relies on the network to implicitly learn various camera models from image appearance, linking image size to real-world dimensions. We believe this approach strains data diversity and network capacity, resulting in lower performance.

相比之下，我们的两种 CSTM 方法都使模型具备了度量预测能力，并取得了相当的性能。表 1 证实了这种相当的性能。因此，在训练过程中调整监督方式和输入图像的外观能有效解决度量模糊问题。此外，我们将我们的方法与 CamConvs [38] 进行了比较，CamConvs [38] 在解码器中使用 4 通道特征融入了相机模型。虽然 CamConvs 采用了相同的训练计划、模型和数据，但它依赖网络从图像外观中隐式学习各种相机模型，将图像尺寸与现实世界的尺寸联系起来。我们认为这种方法会限制数据多样性和网络容量，导致性能较低。

Ablation on canonical space. We investigate the impact of the canonical camera, specifically the canonical focal length. The model is trained on a small sampled dataset and evaluated on both the training and validation sets. We calculate the average Absolute Relative (AbsRel) error for three different focal lengths:250,500,1000,1500, and 2500 . Our experiments indicate that a focal length of 1000 yields slightly better performance than the others; further details can be found in the supplementary materials.

正则空间的消融实验。我们研究了正则相机的影响，特别是正则焦距的影响。模型在一个小的采样数据集上进行训练，并在训练集和验证集上进行评估。我们计算了三种不同焦距 (250、500、1000、1500 和 2500) 下的平均绝对相对误差 (AbsRel)。我们的实验表明，焦距为 1000 时的性能略优于其他焦距；更多细节可在补充材料中找到。

TABLE 9 - Effectiveness of our CSTM. CamConvs [38] directly encodes various camera models in the network, while we perform a simple yet effective transformation to solve the metric ambiguity. Without CSTM, the model achieve transferable metric prediction ability.

表 9 - 我们的 CSTM 的有效性。CamConvs [38] 直接在网络中编码各种相机模型，而我们进行了一种简单而有效的变换来解决度量模糊问题。没有 CSTM 时，模型无法实现可迁移的度量预测能力。

Method	DDAD	Lyft	DS	NS	KITTI	NYU
	Test setof train.data (AbsRel↓)			Zero-shot test set(AbsRel↓)		
w/o CSTM	0.530	0.582	0.394	1.00	0.568	0.584
CamConvs [38]	0.295	0.315	0.213	0.423	0.178	0.333
Ours CSTM_image	0.190	0.235	0.182	0.197	0.097	0.210
Ours CSTM_label	0.183	0.221	0.201	0.213	0.081	0.212

方法	DDAD	来福车 (Lyft)	DS	NS	基蒂数据集 (KITTI)	纽约大学数据集 (NYU)
	训练数据测试集 (绝对相对误差 ↓)			零样本测试集 (绝对相对误差 ↓)		
无 CSTM	0.530	0.582	0.394	1.00	0.568	0.584
相机卷积层 (CamConvs [38])	0.295	0.315	0.213	0.423	0.178	0.333
我们的 CSTM 图像	0.190	0.235	0.182	0.197	0.097	0.210
我们的 CSTM 标签	0.183	0.221	0.201	0.213	0.081	0.212

TABLE 10 - Effectiveness of random proposal normalization loss. Baseline is supervised by ’ $L_{\text{PWN}} + L_{\text{VNL}} + L_{\text{silog}}$ ’. SSIL is the scale-shift invariant loss proposed in [27].

表 10 - 随机提议归一化损失的有效性。基线由“$L_{\text{PWN}} + L_{\text{VNL}} + L_{\text{silog}}$’”监督。SSIL 是文献 [27] 中提出的尺度-平移不变损失 (Scale-Shift Invariant Loss)。

Method	DDAD	Lyft	DS	NS	KITTI	NYUv2
	Test set of train. data (AbsRel↓)				Zero-shot test set	(AbsRel↓)
baseline	0.204	0.251	0.184	0.207	0.104	0.230
baseline + SSIL [27]	0.197	0.263	0.259	0.206	0.105	0.216
baseline + RPNL	0.190	0.235	0.182	0.197	0.097	0.210

方法	DDAD	来福车 (Lyft)	DS	NS	基蒂数据集 (KITTI)	纽约大学深度数据集 v2(NYUv2)
	训练数据测试集 (绝对相对误差 ↓)				零样本测试集	(绝对相对误差 ↓)
基线	0.204	0.251	0.184	0.207	0.104	0.230
基线 + 自监督实例学习 (SSIL) [27]	0.197	0.263	0.259	0.206	0.105	0.216
基线 + 区域提议网络损失 (RPNL)	0.190	0.235	0.182	0.197	0.097	0.210

Effectiveness of the random proposal normalization loss. To demonstrate the effectiveness of our random proposal normalization loss (RPNL), we conducted experiments on a sampled small dataset, with results shown in Table 10. We tested on DDAD, Lyft, DrivingStereo (DS), NuScenes (NS), KITTI, and NYUv2. The ’baseline’

includes all losses except RPNL, which we compare to 'baseline + RPNL' and 'baseline + SSIL [27]'. Our RPNL significantly enhances performance, while the scale-shift invariant loss [27], which normalizes the entire image, offers slight improvements.

随机提议归一化损失的有效性。为了证明我们提出的随机提议归一化损失 (Random Proposal Normalization Loss，RPNL) 的有效性，我们在一个抽样的小数据集上进行了实验，结果如表 10 所示。我们在 DDAD、Lyft、DrivingStereo(DS)、NuScenes(NS)、KITTI 和 NYUv2 数据集上进行了测试。“基线”(baseline) 包含除 RPNL 之外的所有损失，我们将其与“基线 + RPNL”和“基线 + SSIL [27]”进行比较。我们的 RPNL 显著提高了性能，而对整个图像进行归一化的尺度平移不变损失 [27] 仅带来了轻微的改进。

Effectiveness of joint optimization. We assess the impact of joint optimization on both depth and normal estimation using small datasets sampled with ViT-small models over a 4-step iteration. The evaluation is conducted on the NYU indoor dataset and the DIODE outdoor dataset, both of which include normal labels for the convenience of evaluation. In Tab. 11, we start by training the same-architecture networks 'without depth' or 'without normal' prediction. Compared to our joint optimization approach, both single-modality models exhibit slightly worse performance.To further demonstrate the benefit of joint optimization and the incorporation of large-scale outdoor data prior to normal estimation, we train a model using only the Taskonomy dataset (i.e., 'W.o. mixed datasets'), which shows inferior results on DIODE(outdoor). We also verify the effectiveness of the recurrent blocks and the consistency loss. Removing either of them ('W.o. consistency' / 'W.o. recurrent block') could lead to drastic performance degradation for normal estimation, particularly for outdoor scenarios like DIODE(Outdoor). Furthermore, we present some visualization comparisons in Fig 16. Training surface normal and depth together without the consistency loss ('W.o. consistency') results in notably poorer predicted normals compared to our full method ('Ours normal'). Additionally, if the model learns the normal individually ('W.o. depth'), the performance also degrades. The efficiency analysis of the joint optimization module is presented in the supplementary materials.

联合优化的有效性。我们使用通过 ViT-small 模型抽样的小数据集，在 4 步迭代过程中评估联合优化对深度和法线估计的影响。评估在 NYU 室内数据集和 DIODE 室外数据集上进行，这两个数据集都包含法线标签，方便进行评估。在表 11 中，我们首先训练了“无深度”或“无法线”预测的同架构网络。与我们的联合优化方法相比，两种单模态模型的性能都稍差。为了进一步证明联合优化以及在法线估计之前引入大规模室外数据的好处，我们仅使用 Taskonomy 数据集 (即“无混合数据集”) 训练了一个模型，该模型在 DIODE(室外) 数据集上的结果较差。我们还验证了循环块和一致性损失的有效性。移除其中任何一个 (“无一致性”/“无循环块”) 都会导致法线估计性能急剧下降，特别是在像 DIODE(室外) 这样的室外场景中。此外，我们在图 16 中展示了一些可视化比较。在没有一致性损失的情况下同时训练表面法线和深度 (“无一致性”)，与我们的完整方法 (“我们的法线”) 相比，预测的法线明显更差。此外，如果模型单独学习法线 (“无深度”)，性能也会下降。联合优化模块的效率分析在补充材料中给出。

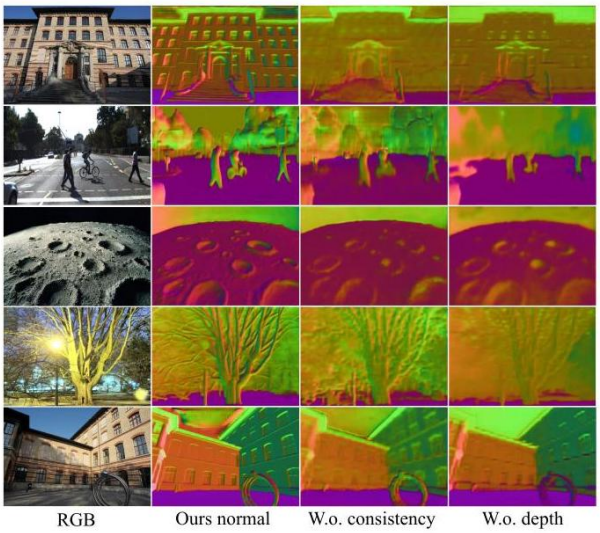

Fig. 16 - Effect of joint depth-normal optimization. We compare normal maps learned by different strategies on several outdoor examples. Learning normal only 'without depth' leads to flattened surfaces, since most of the normal labels lie on planes. In addition, 'without consistency' imposed between depth and normal, the predictions become much coarser.

图 16 - 深度 - 法线联合优化的效果。我们在几个室外示例上比较了不同策略学习到的法线图。仅“无深度”地学习法线会导致表面变平，因为大多数法线标签都位于平面上。此外，在深度和法线之间没有施加“一致性”时，预测结果会变得粗糙得多。

TABLE 11 - Effectiveness of joint optimization. Joint optimization surpasses independent estimation. For outdoor normal estimation, this module introduces geometry clues from large-scale depth data. The proposed recurrent block and depth-normal consistency constraint are essential for the optimization

表 11 - 联合优化的有效性。联合优化优于独立估计。对于室外法线估计，该模块从大规模深度数据中引入几何线索。所提出的循环块和深度 - 法线一致性约束对于优化至关重要。

Method	DIODE(Outdoor) NYU Depth (AbsRel↓)		DIODE(Outdoor) Normal (Med. error↓)	NYU
W.o. normal	0.315	0.119	-	-
W.o. depth	-	-	16.25	8.78
W.o mixed datasets	0.614	0.116	18.94	9.50
W.o. recurrent block	0.309	0.127	16.51	9.31
W.o. consistency	0.310	0.121	16.45	9.72
Ours	0.304	0.114	14.91	8.77

方法	DIODE(户外) 纽约大学深度数据集 (绝对相对误差 ↓)		DIODE(户外) 法线数据集 (中值误差 ↓)	纽约大学 (NYU)
无法线信息	0.315	0.119	-	-
无深度信息	-	-	16.25	8.78
无混合数据集	0.614	0.116	18.94	9.50
无循环块	0.309	0.127	16.51	9.31
无一致性约束	0.310	0.121	16.45	9.72
我们的方法	0.304	0.114	14.91	8.77

Selection of intermediate normal representation. During optimization, unnormalized normal vectors are utilized as the intermediate representation. Here we explore three additional representations (1) A vector defined in so3 representing 3D rotation upon a reference direction. We implement this vector by lietorch [37]. (2) An azimuthal angle and a polar angle. (3) A 2D homogeneous vector [81]. All the representations investigated are additive and can be surjectively transferred into surface normal. In this experiment, we only change the representations and compare the performances. Surprisingly, according to Table 12, the naive unnormalized normal performs the best. We hypothesize that this simplest representation reduces the learning difficulty.

中间法线表示的选择。在优化过程中，未归一化的法线向量被用作中间表示。在这里，我们探索了另外三种表示方式:(1) 一个在 so3(特殊正交群) 中定义的向量，表示相对于参考方向的三维旋转。我们使用 lietorch [37] 来实现这个向量。(2) 方位角和极角。(3) 二维齐次向量 [81]。所有研究的表示方式都是可加的，并且可以满射转换为表面法线。在这个实验中，我们仅改变表示方式并比较性能。令人惊讶的是，根据表 12，简单的未归一化法线表现最佳。我们假设这种最简单的表示方式降低了学习难度。

TABLE 12 - More selection of intermediate normal representation.

表 12 - 更多中间法线表示的选择。

Representation	Taskonomy Scannet Test set (Med. err.↓)		DIODE(Outdoor) Zero-shot testing (Med. err.↓)	NYU
3D rotation vector	5.28	8.92	16.00	9.45
Azi. and polar angles	5.34	9.01	15.21	9.21
Homo. 2D vector [81]	5.02	8.50	15.40	8.79
Ours (unnormalized 3D)	5.01	8.41	14.91	8.77

表示 (Representation)	任务分类法扫描网络测试集 (中值误差 ↓)		DIODE(户外) 零样本测试 (中值误差 ↓)	纽约大学 (NYU)
三维旋转向量	5.28	8.92	16.00	9.45
方位角和极角	5.34	9.01	15.21	9.21
齐次二维向量 [81]	5.02	8.50	15.40	8.79
我们的方法 (未归一化的三维)	5.01	8.41	14.91	8.77

Best optimizing steps To determine the optimal number of optimization steps for various ViT models, we vary different steps to refine depth and normal. Table 13 illustrates that increasing the number of iteration steps does not consistently improve results. Moreover, the ideal number of steps may differ based on the model size, with larger models generally benefiting from more extensive optimization.

最佳优化步骤为了确定各种视觉 Transformer(ViT) 模型的最佳优化步骤数，我们改变不同的步骤来细化深度和法线。表 13 显示，增加迭代步骤数并不总是能改善结果。此外，理想的步骤数可能因模型大小而异，通常较大的模型从更广泛的优化中受益更多。

TABLE 13 - Select the best joint optimizing steps for different ViT models. We find the best step varying with model size. All models are trained following the settings in Tab. 11

表 13 - 为不同的视觉 Transformer(ViT) 模型选择最佳联合优化步骤。我们发现最佳步骤会随模型大小而变化。所有模型均按照表 11 中的设置进行训练

Backbone	ViT-Small	ViT-Large	ViT-giant
#Steps	KITTI Depth (AbsRel↓) / NYU v2 Normal (Med. err.↓)		
2	0.102/9.01	0.070 / 8.40	0.069 / 8.25
4	**0.088**/**8.77**	0.067 / 8.24	0.067 / 8.23
8	0.090 / 8.80	0.065/8.21	0.064/8.22
16	0.095 / 8.79	0.068 / 8.30	0.065 / 8.27

主干网络	小型视觉 Transformer(ViT-Small)	大型视觉 Transformer(ViT-Large)	巨型视觉 Transformer(ViT-giant)
# 步数	KITTI 深度数据集 (绝对相对误差 ↓)/ NYU v2 法线数据集 (中值误差 ↓)		
2	0.102/9.01	0.070 / 8.40	0.069 / 8.25
4	**0.088**/**8.77**	0.067 / 8.24	0.067 / 8.23
8	0.090 / 8.80	0.065/8.21	0.064/8.22
16	0.095 / 8.79	0.068 / 8.30	0.065 / 8.27

5 CONCLUSION

5 结论

In this paper, we introduce a family of geometric foundation models for zero-shot monocular metric depth and surface normal estimation. We propose solutions to address challenges in both metric depth estimation and surface normal estimation. To resolve depth ambiguity caused by varying focal lengths, we present a novel canonical camera space transformation method. Additionally, to overcome the scarcity of outdoor normal data labels, we introduce a joint depth-normal optimization framework that leverages knowledge from large-scale depth annotations.

在本文中，我们介绍了一系列用于零样本单目度量深度和表面法线估计的几何基础模型。我们针对度量深度估计和表面法线估计中的挑战提出了解决方案。为解决因焦距变化导致的深度模糊问题，我们提出了一种新颖的规范相机空间变换方法。此外，为克服户外法线数据标签稀缺的问题，我们引入了一个联合深度 - 法线优化框架，该框架利用了大规模深度标注的知识。

Our approach enables the integration of millions of data samples captured by over 10,000 cameras to train a unified metric-depth and surface-normal model. To enhance the model's robustness, we curate a dataset comprising over 16 million samples for training. Zero-shot evaluations demonstrate the effectiveness and robustness of our method. For downstream applications, our models are capable of reconstructing metric 3D from a single view, enabling metrology on randomly collected internet images and dense mapping of large-scale scenes. With their precision, generalization, and versatility, Metric3D v2 models serve as geometric foundational models for monocular perception.

我们的方法能够整合由一万多台相机捕获的数百万个数据样本，以训练一个统一的度量深度和表面法线模型。为增强模型的鲁棒性，我们精心整理了一个包含超过一千六百万个样本的数据集用于训练。零样本评估证明了我们方法的有效性和鲁棒性。对于下游应用，我们的模型能够从单视图重建度量三维信息，从而能够对随机收集的互联网图像进行测量，并对大规模场景进行密集映射。凭借其精度、泛化能力和多功能性，Metric3D v2 模型可作为单目感知的几何基础模型。

REFERENCES

参考文献

[1] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen, ”Dense 3d object reconstruction from a single depth view,” IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 12, pp. 2820-2834, 2018. 1

B. Yang、S. Rosa、A. Markham、N. Trigoni 和 H. Wen，“从单深度视图进行密集三维物体重建”,《IEEE 模式分析与机器智能汇刊》，第 41 卷，第 12 期，第 2820 - 2834 页，2018 年。1

[2] J. Ju, C. W. Tseng, O. Bailo, G. Dikov, and M. Ghafoorian, ”Dg-recon: Depth-guided neural 3d scene reconstruction,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 18184-18194, 2023. 1

J. Ju、C. W. Tseng、O. Bailo、G. Dikov 和 M. Ghafoorian，“Dg - recon: 深度引导的神经三维场景重建”,《IEEE/CVF 国际计算机视觉会议论文集》，第 18184 - 18194 页，2023 年。1

[3] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, ”Occupancy networks: Learning 3d reconstruction in function space,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 4460-4470, 2019. 1

L. Mescheder、M. Oechsle、M. Niemeyer、S. Nowozin 和 A. Geiger，“占用网络: 在函数空间中学习三维重建”,《IEEE 计算机视觉与模式识别会议论文集》，第 4460 - 4470 页，2019 年。1

[4] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, ”Depth-supervised nerf: Fewer views and faster training for free,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12882-12891, 2022. 1

K. Deng、A. Liu、J. - Y. Zhu 和 D. Ramanan，“深度监督的神经辐射场: 免费实现更少视图和更快训练”,《IEEE/CVF 计算机视觉与模式识别会议论文集》，第 12882 - 12891 页，2022 年。1

[5] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, and M. Nießner, ”Dense depth priors for neural radiance fields from sparse input views,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12892-12901, 2022. 1

B. Roessle、J. T. Barron、B. Mildenhall、P. P. Srinivasan 和 M. Nießner，“从稀疏输入视图为神经辐射场提供密集深度先验”,《IEEE/CVF 计算机视觉与模式识别会议论文集》，第 12892 - 12901 页，2022 年。1

[6] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, ”Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction,” Advances in neural information processing systems, vol. 35, pp. 25018-25032, 2022. 1, 5

Z. Yu、S. Peng、M. Niemeyer、T. Sattler 和 A. Geiger，“Monosdf: 探索单目几何线索用于神经隐式表面重建”,《神经信息处理系统进展》，第 35 卷，第 25018 - 25032 页，2022 年。1, 5

[7] C. Jiang, H. Zhang, P. Liu, Z. Yu, H. Cheng, B. Zhou, and S. Shen, ”H2-mapping: Real-time dense mapping using hierarchical hybrid representation,” arXiv preprint arXiv:2306.03207, 2023. 1

C. Jiang、H. Zhang、P. Liu、Z. Yu、H. Cheng、B. Zhou 和 S. Shen，“H2 - mapping: 使用分层混合表示进行实时密集映射”，预印本 arXiv:2306.03207，2023 年。1

[8] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, ”Bevdepth: Acquisition of reliable depth for multi-view 3d object detection,” arXiv: Comp. Res. Repository, p. 2206.10092, 2022. 1

Y. Li、Z. Ge、G. Yu、J. Yang、Z. Wang、Y. Shi、J. Sun 和 Z. Li，“Bevdepth: 为多视图三维物体检测获取可靠深度”，预印本 arXiv:2206.10092，2022 年。1

[9] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, ”Fb-occ: 3d occupancy prediction based on forward-backward view transformation,” arXiv preprint arXiv:2307.01492, 2023. 1

Z. Li、Z. Yu、D. Austin、M. Fang、S. Lan、J. Kautz 和 J. M. Alvarez，“Fb - occ: 基于前后视图变换的三维占用预测”，预印本 arXiv:2307.01492，2023 年。1

[10] R. Fan, H. Wang, P. Cai, and M. Liu, ”Sne-roadseg: Incorporat-

R. Fan、H. Wang、P. Cai 和 M. Liu，“Sne - roadseg: 将

ing surface normal information into semantic segmentation for accurate freespace detection,” in European Conference on Computer Vision, pp. 340-356, Springer, 2020. 1

表面法线信息融入语义分割以实现准确的自由空间检测”，《欧洲计算机视觉会议论文集》，第 340 - 356 页，施普林格出版社，2020 年。1

[11] J. Behley and C. Stachniss, ”Efficient surfel-based slam using 3d laser range data in urban environments.,” in Robotics: Science and Systems, vol. 2018, p. 59, 2018. 1, 5

J. 贝赫利 (J. Behley) 和 C. 施塔希尼斯 (C. Stachniss)，“城市环境中基于高效曲面元素 (surfel) 的三维激光测距数据同步定位与地图构建 (SLAM)”，载于《机器人科学与系统》(Robotics: Science and Systems)，2018 年第 2018 卷，第 59 页，2018 年。1, 5

[12] R. Mur-Artal and J. D. Tardós, ”ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras,” IEEE Trans. Robot., vol. 33, no. 5, pp. 1255-1262, 2017. 1, 13

R. 穆尔 - 阿塔尔 (R. Mur - Artal) 和 J. D. 塔尔多斯 (J. D. Tardós)，“ORB - SLAM2: 用于单目、双目和 RGB - D 相机的开源 SLAM 系统”，《电气与电子工程师协会机器人汇刊》(IEEE Trans. Robot.)，第 33 卷，第 5 期，第 1255 - 1262 页，2017 年。1, 13

[13] T. Schops, T. Sattler, and M. Pollefeys, ”Bad slam: Bundle adjusted direct rgb-d slam,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 134-144, 2019. 1

T. 肖普斯 (T. Schops)、T. 萨特勒 (T. Sattler) 和 M. 波勒菲斯 (M. Pollefeys)，“Bad SLAM: 捆绑调整的直接 RGB - D SLAM”，载于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》(Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition)，第 134 - 144 页，2019 年。1

[14] H. Zhu, H. Yang, X. Wu, D. Huang, S. Zhang, X. He, T. He, H. Zhao, C. Shen, Y. Qiao, et al., ”Ponderv2: Pave the way for 3 d foundataion model with a universal pre-training paradigm,” arXiv preprint arXiv:2310.08586, 2023. 1

朱浩 (H. Zhu)、杨浩 (H. Yang)、吴翔 (X. Wu)、黄迪 (D. Huang)、张帅 (S. Zhang)、何鑫 (X. He)、何涛 (T. He)、赵辉 (H. Zhao)、沈超 (C. Shen)、乔宇 (Y. Qiao) 等，“Ponderv2: 以通用预训练范式为 3 d 基础模型铺平道路”，预印本 arXiv:2310.08586，2023 年。1

[15] J. Zhou, J. Wang, B. Ma, Y.-S. Liu, T. Huang, and X. Wang, ”Uni3d: Exploring unified 3d representation at scale,” arXiv preprint arXiv:2310.06773, 2023. 1

周杰 (J. Zhou)、王军 (J. Wang)、马博 (B. Ma)、刘一杉 (Y. - S. Liu)、黄涛 (T. Huang) 和王鑫 (X. Wang)，“Uni3d: 大规模探索统一三维表示”，预印本 arXiv:2310.06773，2023 年。1

[16] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, ”Unifying flow, stereo and depth estimation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023. 1

徐航 (H. Xu)、张军 (J. Zhang)、蔡杰 (J. Cai)、H. 雷扎托菲吉 (H. Rezatofighi)、余峰 (F. Yu)、陶大程 (D. Tao) 和 A. 盖格 (A. Geiger)，“统一光流、立体视觉和深度估计”，《电气与电子工程师协会模式分析与机器智能汇刊》(IEEE Transactions on Pattern Analysis and Machine Intelligence)，2023 年。1

[17] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, ”New CRFs: Neural window fully-connected CRFs for monocular depth estimation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2022. 1, 2, 4, 9, 10

袁伟 (W. Yuan)、顾鑫 (X. Gu)、戴泽 (Z. Dai)、朱帅 (S. Zhu) 和谭平 (P. Tan)，“新型条件随机场 (CRFs): 用于单目深度估计的神经窗口全连接 CRFs”，载于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，2022 年。1, 2, 4, 9, 10

[18] W. Yin, Y. Liu, and C. Shen, ”Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction,” IEEE Trans. Pattern Anal. Mach. Intell., 2021. 1, 2, 4, 5, 8, 10

尹伟 (W. Yin)、刘阳 (Y. Liu) 和沈超 (C. Shen)，“虚拟法线: 为准确稳健的深度预测施加几何约束”，《电气与电子工程师协会模式分析与机器智能汇刊》(IEEE Trans. Pattern Anal. Mach. Intell.)，2021 年。1, 2, 4, 5, 8, 10

[19] S. F. Bhat, I. Alhashim, and P. Wonka, ”Adabins: Depth estimation using adaptive bins,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 4009-4018, 2021. 1, 2, 4, 9, 10

S. F. 巴特 (S. F. Bhat)、I. 阿尔哈希姆 (I. Alhashim) 和 P. 翁卡 (P. Wonka)，“自适应分箱 (Adabins): 使用自适应分箱进行深度估计”，载于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 4009 - 4018 页，2021 年。1, 2, 4, 9, 10

[20] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, ”Transformer-based attention networks for continuous pixel-wise prediction,” in Proc. IEEE Int. Conf. Comp. Vis., 2021. 1, 2, 9

杨刚 (G. Yang)、唐浩 (H. Tang)、丁明 (M. Ding)、N. 塞贝 (N. Sebe) 和 E. 里奇 (E. Ricci)，“基于 Transformer 的注意力网络用于连续逐像素预测”，载于《电气与电子工程师协会国际计算机视觉会议论文集》(Proc. IEEE Int. Conf. Comp. Vis.)，2021 年。1, 2, 9

[21] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, ”Monocular relative depth perception with web stereo data supervision,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 311- 320,2018.2,5

冼康 (K. Xian)、沈超 (C. Shen)、曹政 (Z. Cao)、卢航 (H. Lu)、肖扬 (Y. Xiao)、李锐 (R. Li) 和罗泽 (Z. Luo)，“基于网络立体数据监督的单目相对深度感知”，载于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 311 - 320 页，2018 年。2, 5

[22] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, ”Structure-guided ranking loss for single image depth prediction,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 611-620, 2020. 2, 5

冼康 (K. Xian)、张军 (J. Zhang)、王奥 (O. Wang)、麦磊 (L. Mai)、林泽 (Z. Lin) 和曹政 (Z. Cao)，“用于单图像深度预测的结构引导排序损失”，载于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 611 - 620 页，2020 年。2, 5

[23] W. Chen, S. Qian, D. Fan, N. Kojima, M. Hamilton, and J. Deng, ”Oasis: A large-scale dataset for single image 3 d in the wild,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 679-688, 2020. 2, 5

陈巍 (W. Chen)、钱森 (S. Qian)、范迪 (D. Fan)、小岛直树 (N. Kojima)、M. 汉密尔顿 (M. Hamilton) 和邓军 (J. Deng)，“Oasis: 用于野外单图像 3 d 的大规模数据集”，载于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 679 - 688 页，2020 年。2, 5

[24] W. Chen, Z. Fu, D. Yang, and J. Deng, ”Single-image depth perception in the wild,” in Proc. Advances in Neural Inf. Process. Syst., pp. 730-738, 2016. 2, 5

陈巍 (W. Chen)、付泽 (Z. Fu)、杨迪 (D. Yang) 和邓军 (J. Deng)，“野外单图像深度感知”，载于《神经信息处理系统进展会议论文集》(Proc. Advances in Neural Inf. Process. Syst.)，第 730 - 738 页，2016 年。2, 5

[25] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, and C. Shen, ”Learning to recover 3d scene shape from a single image,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2021. 2, 3, 4,5,8,9,10,12,13

尹伟 (W. Yin)、张军 (J. Zhang)、王奥 (O. Wang)、尼克劳斯 (S. Niklaus)、麦磊 (L. Mai)、陈硕 (S. Chen) 和沈超 (C. Shen)，“从单张图像中学习恢复三维场景形状”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，2021 年。2, 3, 4,5,8,9,10,12,13

[26] W. Yin, J. Zhang, O. Wang, S. Niklaus, S. Chen, Y. Liu, and C. Shen, ”Towards accurate reconstruction of 3d scene shape from a single monocular image,” IEEE Trans. Pattern Anal. Mach. Intell., 2022. 2, 5

尹伟 (W. Yin)、张军 (J. Zhang)、王奥 (O. Wang)、尼克劳斯 (S. Niklaus)、陈硕 (S. Chen)、刘阳 (Y. Liu) 和沈超 (C. Shen)，“从单目图像实现三维场景形状的精确重建”，《电气与电子工程师协会模式分析与机器智能汇刊》(IEEE Trans. Pattern Anal. Mach. Intell.)，2022 年。2, 5

[27] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, ”Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” IEEE Trans. Pattern Anal. Mach. Intell., 2020. 2, 3, 5, 8, 10, 14

兰夫特尔 (R. Ranftl)、拉辛格 (K. Lasinger)、哈夫纳 (D. Hafner)、辛德勒 (K. Schindler) 和科尔图恩 (V. Koltun)，“迈向鲁棒的单目深度估计: 混合数据集以实现零样本跨数据集迁移”，《电气与电子工程师协会模式分析与机器智能汇刊》(IEEE Trans. Pattern Anal. Mach. Intell.)，2020 年。2, 3, 5, 8, 10, 14

[28] R. Ranftl, A. Bochkovskiy, and V. Koltun, ”Vision transformers for dense prediction,” in Proc. IEEE Int. Conf. Comp. Vis., pp. 12179- 12188,2021.2,8,10,12,13

兰夫特尔 (R. Ranftl)、博奇科夫斯基 (A. Bochkovskiy) 和科尔图恩 (V. Koltun)，“用于密集预测的视觉 Transformer”，收录于《电气与电子工程师协会国际计算机视觉会议论文集》(Proc. IEEE Int. Conf. Comp. Vis.)，第 12179 - 12188 页，2021 年。2,8,10,12,13

[29] C. Zhang, W. Yin, Z. Wang, G. Yu, B. Fu, and C. Shen, ”Hierarchical normalization for robust monocular depth estimation,” Proc. Advances in Neural Inf. Process. Syst., 2022. 2, 3, 5, 9, 10

张晨 (C. Zhang)、尹伟 (W. Yin)、王哲 (Z. Wang)、余刚 (G. Yu)、傅博 (B. Fu) 和沈超 (C. Shen)，“用于鲁棒单目深度估计的分层归一化”，《神经信息处理系统进展会议论文集》(Proc. Advances in Neural Inf. Process. Syst.)，2022 年。2, 3, 5, 9, 10

[30] W. Yin, Y. Liu, C. Shen, and Y. Yan, ”Enforcing geometric constraints of virtual normal for depth prediction,” in Proc. IEEE Int. Conf. Comp. Vis., 2019. 2, 4, 9

尹伟 (W. Yin)、刘阳 (Y. Liu)、沈超 (C. Shen) 和闫宇 (Y. Yan)，“在深度预测中施加虚拟法线的几何约束”，收录于《电气与电子工程师协会国际计算机视觉会议论文集》(Proc. IEEE Int. Conf. Comp. Vis.)，2019 年。2, 4, 9

[31] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, ”Repurposing diffusion-based image generators for

柯斌 (B. Ke)、奥布霍夫 (A. Obukhov)、黄硕 (S. Huang)、梅茨格 (N. Metzger)、多 (R. C. Daudt) 和辛德勒 (K. Schindler)，“将基于扩散的图像生成器重新用于

monocular depth estimation,” arXiv preprint arXiv:2312.02145, 2023.2

单目深度估计”，预印本 arXiv:2312.02145，2023 年。2

[32] T. Do, K. Vuong, S. I. Roumeliotis, and H. S. Park, ”Surface normal estimation of tilted images via spatial rectifier,” in Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV 16, pp. 265-280, Springer, 2020.

杜涛 (T. Do)、武勇 (K. Vuong)、鲁梅利奥蒂斯 (S. I. Roumeliotis) 和朴河善 (H. S. Park)，“通过空间矫正器对倾斜图像进行表面法线估计”，收录于《计算机视觉 - 欧洲计算机视觉会议 2020: 第 16 届欧洲会议，英国格拉斯哥，2020 年 8 月 23 - 28 日，会议录，第四部分 16》(Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23 - 28, 2020, Proceedings, Part IV 16)，第 265 - 280 页，施普林格出版社，2020 年。

2, 5, 9

[33] G. Bae, I. Budvytis, and R. Cipolla, ”Estimating and exploiting the aleatoric uncertainty in surface normal estimation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13137-13146, 2021. 2, 5, 8, 9, 10, 11, 12

裴圭 (G. Bae)、布德维蒂斯 (I. Budvytis) 和奇波拉 (R. Cipolla)，“估计并利用表面法线估计中的随机不确定性”，收录于《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集》(Proceedings of the IEEE/CVF International Conference on Computer Vision)，第 13137 - 13146 页，2021 年。2, 5, 8, 9, 10, 11, 12

[34] X. Yang, L. Yuan, K. Wilber, A. Sharma, X. Gu, S. Qiao, S. Debats, H. Wang, H. Adam, M. Sirotenko, et al., ”Polymax: General dense prediction with mask transformer,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1050- 1061, 2024. 2, 5, 9, 10

杨晓 (X. Yang)、袁磊 (L. Yuan)、威尔伯 (K. Wilber)、夏尔马 (A. Sharma)、顾翔 (X. Gu)、乔硕 (S. Qiao)、德巴茨 (S. Debats)、王浩 (H. Wang)、亚当 (H. Adam)、西罗坚科 (M. Sirotenko) 等，“Polymax: 使用掩码 Transformer 进行通用密集预测”，收录于《电气与电子工程师协会/计算机视觉基金会冬季计算机视觉应用会议论文集》(Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision)，第 1050 - 1061 页，2024 年。2, 5, 9, 10

[35] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, ”Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 10786- 10796, 2021. 2, 3, 5, 9, 10, 11, 12

埃夫特哈尔 (A. Eftekhar)、萨克斯 (A. Sax)、马利克 (J. Malik) 和扎米尔 (A. Zamir)，“Omnidata: 一种从三维扫描创建多任务中级视觉数据集的可扩展管道”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 10786 - 10796 页，2021 年。2, 3, 5, 9, 10, 11, 12

[36] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, ”Zoedepth: Zero-shot transfer by combining relative and metric depth,” arXiv preprint arXiv:2302.12288, 2023. 3, 5, 9, 10, 11, 12

巴特 (S. F. Bhat)、比尔克尔 (R. Birkl)、沃夫克 (D. Wofk)、翁卡 (P. Wonka) 和米勒 (M. Müller),“Zoedepth: 通过结合相对深度和度量深度实现零样本迁移”，预印本 arXiv:2302.12288，2023 年。3, 5, 9, 10, 11, 12

[37] Z. Teed and J. Deng, ”Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras,” vol. 34, pp. 16558-16569, 2021. 3, 4,12,13,15

泰德 (Z. Teed) 和邓军 (J. Deng),“Droid - SLAM: 用于单目、立体和 RGB - D 相机的深度视觉 SLAM”，第 34 卷，第 16558 - 16569 页，2021 年。3, 4,12,13,15

[38] J. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, ”CAM-Convs: camera-aware multi-scale convolutions for single-view depth,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 11826-11835, 2019. 2, 14

J. 法西尔 (J. Facil)、B. 乌门霍费尔 (B. Ummenhofer)、H. 周 (H. Zhou)、L. 蒙特萨诺 (L. Montesano)、T. 布罗克斯 (T. Brox) 和 J. 西韦拉 (J. Civera)，“CAM-Convs: 用于单视图深度的相机感知多尺度卷积”，载于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 11826 - 11835 页，2019 年。2，14

[39] S. Peng, S. Zhang, Z. Xu, C. Geng, B. Jiang, H. Bao, and X. Zhou, ”Animatable neural implicit surfaces for creating avatars from videos,” arXiv: Comp. Res. Repository, p. 2203.08133, 2022. 3

彭 (Peng)、张 (Zhang)、徐 (Xu)、耿 (Geng)、姜 (Jiang)、鲍 (Bao) 和周 (Zhou)，“用于从视频创建虚拟形象的可动画神经隐式曲面”，预印本服务器 (arXiv): 计算机研究库，第 2203.08133 页，2022 年。3

[40] J. Huang, Y. Zhou, T. Funkhouser, and L. J. Guibas, ”Framenet: Learning local canonical frames of 3 d surfaces from a single rgb image,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8638-8647, 2019. 3, 9, 10

黄 (Huang)、周 (Zhou)、芬克豪泽 (Funkhouser) 和吉巴斯 (Guibas)，“框架网络: 从单张 RGB 图像中学习 3 d 曲面的局部规范框架”，载于《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集》，第 8638 - 8647 页，2019 年。3, 9, 10

[41] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, ”Geonet: Geometric neural network for joint depth and surface normal estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 283-291, 2018. 3, 5, 8, 9

齐 X.(X. Qi)、廖 R.(R. Liao)、刘 Z.(Z. Liu)、乌尔塔松 R.(R. Urtasun) 和贾 J.(J. Jia)，“Geonet: 用于联合深度和表面法线估计的几何神经网络”，载于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 283 - 291 页，2018 年。3, 5, 8, 9

[42] G. Bae, I. Budvytis, and R. Cipolla, ”Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty,” arXiv preprint arXiv:2210.03676, 2022.3, 5, 7, 9, 10

G. 裴 (Bae)、I. 布德维蒂斯 (Budvytis) 和 R. 奇波拉 (Cipolla)，“铁深度 (IronDepth): 利用表面法线及其不确定性对单视图深度进行迭代细化”，预印本 arXiv:2210.03676，2022 年 3 月、5 月、7 月、9 月、10 月

[43] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. So Kweon, ”Non-local spatial propagation network for depth completion,” in Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII 16, pp. 120-136, Springer, 2020. 3,5

J. 帕克 (J. Park)、K. 朱 (K. Joo)、Z. 胡 (Z. Hu)、C.-K. 刘 (C.-K. Liu) 和 I. 苏权 (I. So Kweon)，《用于深度补全的非局部空间传播网络》，载于《计算机视觉——ECCV 2020: 第 16 届欧洲会议，英国格拉斯哥，2020 年 8 月 23 - 28 日，会议录，第十三部分 16》，第 120 - 136 页，施普林格出版社，2020 年。3,5

[44] S. Shao, Z. Pei, X. Wu, Z. Liu, W. Chen, and Z. Li, ”Iebins: Iterative elastic bins for monocular depth estimation,” arXiv preprint arXiv:2309.14137, 2023. 3, 5, 7, 9

邵 (Shao)、裴 (Pei)、吴 (Wu)、刘 (Liu)、陈 (Chen) 和李 (Li)，“Iebins: 用于单目深度估计的迭代弹性区间 (Iebins: Iterative elastic bins for monocular depth estimation)”，预印本 arXiv:2309.14137，2023 年。3、5、7、9

[45] L. Lipson, Z. Teed, and J. Deng, ”Raft-stereo: Multilevel recurrent field transforms for stereo matching,” in Int. Conf. 3D. Vis., 2021. 3, 5, 7

L. 利普森 (Lipson)、Z. 蒂德 (Teed) 和 J. 邓 (Deng)，“筏式立体匹配: 用于立体匹配的多级循环场变换”，载于《国际三维视觉会议论文集》(Int. Conf. 3D. Vis.)，2021 年。3, 5, 7

[46] Z. Teed and J. Deng, ”Raft: Recurrent all-pairs field transforms for optical flow,” in Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II 16, pp. 402-419, Springer, 2020. 3, 5, 7

Z. 蒂德 (Z. Teed) 和 J. 邓 (J. Deng)，《Raft: 用于光流的循环全对场变换》，载于《计算机视觉——ECCV 2020: 第 16 届欧洲会议，英国格拉斯哥，2020 年 8 月 23 - 28 日，会议录，第二部分 16》，第 402 - 419 页，施普林格出版社 (Springer)，2020 年。3, 5, 7

[47] L. Sun, W. Yin, E. Xie, Z. Li, C. Sun, and C. Shen, ”Improving monocular visual odometry using learned depth,” IEEE Transactions on Robotics, vol. 38, no. 5, pp. 3173-3186, 2022. 4

孙 (Sun)、尹 (Yin)、谢 (Xie)、李 (Li)、孙 (Sun) 和沈 (Shen)，“利用学习到的深度改进单目视觉里程计”，《IEEE 机器人学汇刊》(IEEE Transactions on Robotics)，第 38 卷，第 5 期，第 3173 - 3186 页，2022 年。4

[48] S. Im, H.-G. Jeon, S. Lin, and I.-S. Kweon, ”Dpsnet: End-to-end deep plane sweep stereo,” in Proc. Int. Conf. Learn. Representations, 2019.4,12,13

林 (S. Im)、全 (H.-G. Jeon)、林 (S. Lin) 和权 (I.-S. Kweon)，“Dpsnet: 端到端深度平面扫描立体视觉 (Dpsnet: End-to-end deep plane sweep stereo)”，收录于《国际学习表征会议论文集》(Proc. Int. Conf. Learn. Representations)，2019 年。4,12,13

[49] R. Mur-Artal and J. D. Tardós, ”Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras,” IEEE transactions on robotics, vol. 33, no. 5, pp. 1255-1262, 2017. 4

R. 穆尔 - 阿尔塔尔 (R. Mur-Artal) 和 J. D. 塔尔多斯 (J. D. Tardós)，“Orb-slam2: 用于单目、立体和 RGB - D 相机的开源 SLAM 系统”，《IEEE 机器人学汇刊》(IEEE transactions on robotics)，第 33 卷，第 5 期，第 1255 - 1262 页，2017 年。4

[50] R. Zhu, X. Yang, Y. Hold-Geoffroy, F. Perazzi, J. Eisenmann, K. Sunkavalli, and M. Chandraker, ”Single view metrology in the wild,” in Proc. Eur. Conf. Comp. Vis., pp. 316-333, Springer, 2020.4

朱 (Zhu)、杨 (Yang)、霍尔德 - 杰弗里 (Hold - Geoffroy)、佩拉齐 (Perazzi)、艾森曼 (Eisenmann)、松卡瓦利 (Sunkavalli) 和钱德拉克尔 (Chandraker)，“野外单视图测量学”，载于《欧洲计算机视觉会议论文集》，第 316 - 333 页，施普林格出版社，2020 年 4 月。

[51] J. T. Barron and J. Malik, ”Shape, illumination, and reflectance from shading,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 8, pp. 1670-1687, 2014. 4

J. T. 巴伦 (Barron) 和 J. 马利克 (Malik)，“从阴影中获取形状、光照和反射率”，《电气与电子工程师协会模式分析与机器智能汇刊》(IEEE Trans. Pattern Anal. Mach. Intell.)，第 37 卷，第 8 期，第 1670 - 1687 页，2014 年。4

[52] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, ”Pixel2mesh: Generating 3d mesh models from single RGB images,” in Proc. Eur. Conf. Comp. Vis., pp. 52-67, 2018. 4

王 (Wang)、张 (Zhang)、李 (Li)、傅 (Fu)、刘 (Liu) 和蒋 (Jiang)，“Pixel2mesh: 从单张 RGB 图像生成 3D 网格模型”，《欧洲计算机视觉会议论文集》(Proc. Eur. Conf. Comp. Vis.)，第 52 - 67 页，2018 年。4

[53] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. Freeman, and J. Tenenbaum,”Learning shape priors for single-view 3d completion and

吴 (Wu)、张 (Zhang)、张 (Zhang)、张 (Zhang)、弗里曼 (Freeman) 和特南鲍姆 (Tenenbaum)，“学习单视图 3D 补全和

reconstruction,” in Proc. Eur. Conf. Comp. Vis., pp. 646-662, 2018. 4

重建的形状先验”，《欧洲计算机视觉会议论文集》(Proc. Eur. Conf. Comp. Vis.)，第 646 - 662 页，2018 年。4

[54] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, ”Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 2304-2314, 2019. 4

斋藤 (Saito)、黄 (Huang)、夏目 (Natsume)、森岛 (Morishima)、金泽 (Kanazawa) 和李 (Li)，“Pifu: 用于高分辨率着装人体数字化的像素对齐隐式函数”，《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 2304 - 2314 页，2019 年。4

[55] S. Saito, T. Simon, J. Saragih, and H. Joo, ”Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 84- 93, 2020. 4

斋藤 (Saito)、西蒙 (Simon)、萨拉吉 (Saragih) 和朱 (Joo)，“Pifuhd: 用于高分辨率 3D 人体数字化的多级像素对齐隐式函数”，《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 84 - 93 页，2020 年。4

[56] A. Saxena, M. Sun, and A. Y. Ng, ”Make3d: Learning 3d scene structure from a single still image,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 5, pp. 824-840, 2008. 4

萨克塞纳 (Saxena)、孙 (Sun) 和吴恩达 (Ng)，“Make3d: 从单张静态图像学习 3D 场景结构”，《电气与电子工程师协会模式分析与机器智能汇刊》(IEEE Trans. Pattern Anal. Mach. Intell.)，第 31 卷，第 5 期，第 824 - 840 页，2008 年。4

[57] C. Zhang, W. Yin, G. Yu, Z. Wang, T. Chen, B. Fu, J. T. Zhou, and C. Shen, ”Robust geometry-preserving depth estimation using differentiable rendering,” in Proc. IEEE Int. Conf. Comp. Vis., 2023. 4

张 (Zhang)、尹 (Yin)、于 (Yu)、王 (Wang)、陈 (Chen)、傅 (Fu)、周 (Zhou) 和沈 (Shen)，“使用可微渲染进行鲁棒的几何保持深度估计”，《电气与电子工程师协会国际计算机视觉会议论文集》(Proc. IEEE Int. Conf. Comp. Vis.)，2023 年。4

[58] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, ”Indoor segmentation and support inference from rgbd images,” in Proc. Eur. Conf. Comp. Vis., pp. 746-760, Springer, 2012. 4, 5, 9, 10

西尔伯曼 (Silberman)、霍耶姆 (Hoiem)、科利 (Kohli) 和弗格斯 (Fergus)，“从 RGB - D 图像进行室内分割和支撑推断”，《欧洲计算机视觉会议论文集》(Proc. Eur. Conf. Comp. Vis.)，第 746 - 760 页，施普林格出版社 (Springer)，2012 年。4、5、9、10

[59] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, ”Vision meets robotics: The kitti dataset,” Int. J. Robot. Res., 2013. 4, 9, 10

盖格 (Geiger)、伦茨 (Lenz)、施蒂勒 (Stiller) 和乌尔塔松 (Urtasun)，“视觉与机器人技术的结合: 基蒂数据集 (Kitti 数据集)”，《国际机器人研究杂志》(Int. J. Robot. Res.)，2013 年。4、9、10

[60] D. Eigen, C. Puhrsch, and R. Fergus, ”Depth map prediction from a single image using a multi-scale deep network,” in Proc. Advances in Neural Inf. Process. Syst., pp. 2366-2374, 2014. 4, 8

艾根 (Eigen)、普尔施 (Puhrsch) 和弗格斯 (Fergus)，“使用多尺度深度网络从单张图像进行深度图预测”，《神经信息处理系统进展会议论文集》(Proc. Advances in Neural Inf. Process. Syst.)，第 2366 - 2374 页，2014 年。4、8

[61] V. Guizilini, I. Vasiljevic, D. Chen, R. Ambrus, and A. Gaidon, ”Towards zero-shot scale-aware monocular depth estimation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9233-9243, 2023. 5, 9, 10

吉齐利尼 (Guizilini)、瓦西列维奇 (Vasiljevic)、陈 (Chen)、安布鲁斯 (Ambrus) 和盖登 (Gaidon)，“迈向零样本尺度感知单目深度估计”，《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集》(Proceedings of the IEEE/CVF International Conference on Computer Vision)，第 9233 - 9243 页，2023 年。5、9、10

[62] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, ”Unidepth: Universal monocular metric depth estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10106-10116, 2024. 5

皮奇内利 (Piccinelli)、杨 (Yang)、萨卡里迪斯 (Sakaridis)、塞古 (Segu)、李 (Li)、范古尔 (Van Gool) 和余 (Yu)，“Unidepth: 通用单目度量深度估计”，《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》(Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition)，第 10106 - 10116 页，2024 年。5

[63] K. Wang, F. Gao, and S. Shen, ”Real-time scalable dense surfel mapping,” in 2019 International conference on robotics and automation (ICRA), pp. 6919-6925, IEEE, 2019. 5

王 (Wang)、高 (Gao) 和沈 (Shen)，“实时可扩展的密集曲面元映射”，《2019 年国际机器人与自动化会议论文集》(2019 International conference on robotics and automation (ICRA))，第 6919 - 6925 页，电气与电子工程师协会 (IEEE)，2019 年。5

[64] J. Wang, P. Wang, X. Long, C. Theobalt, T. Komura, L. Liu, and W. Wang, ”Neuris: Neural reconstruction of indoor scenes using normal priors,” in European Conference on Computer Vision, pp. 139-155, Springer, 2022. 5

王 (Wang)、王 (Wang)、龙 (Long)、特奥博尔特 (Theobalt)、小村 (Komura)、刘 (Liu) 和王 (Wang)，“Neuris: 使用法线先验进行室内场景的神经重建”，《欧洲计算机视觉会议论文集》(European Conference on Computer Vision)，第 139 - 155 页，施普林格出版社 (Springer)，2022 年。5

[65] X. Qi, Z. Liu, R. Liao, P. H. Torr, R. Urtasun, and J. Jia, ”Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 2, pp. 969- 984, 2020. 5, 9

齐 X.、刘 Z.、廖 R.、托尔 P. H.、乌尔塔松 R. 和贾 J.，《Geonet++: 用于联合深度和表面法线估计的具有边缘感知细化的迭代几何神经网络》，《电气与电子工程师协会模式分析与机器智能汇刊》，第 44 卷，第 2 期，第 969 - 984 页，2020 年。5, 9

[66] X. Wang, D. Fouhey, and A. Gupta, ”Designing deep networks for surface normal estimation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 539-547, 2015. 5, 9

王 X.、福伊 D. 和古普塔 A.，《设计用于表面法线估计的深度网络》，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 539 - 547 页，2015 年。5, 9

[67] D. Eigen and R. Fergus, ”Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in Proceedings of the IEEE international conference on computer vision, pp. 2650-2658,2015. 5,9

艾根 D. 和弗格斯 R.，《使用通用多尺度卷积架构预测深度、表面法线和语义标签》，收录于《电气与电子工程师协会国际计算机视觉会议论文集》，第 2650 - 2658 页，2015 年。5, 9

[68] S. Liao, E. Gavves, and C. G. Snoek, ”Spherical regression: Learning viewpoints, surface normals and 3d rotations on n-spheres,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9759-9767, 2019. 5

廖 S.、加维斯 E. 和斯诺克 C. G.，《球面回归: 在 n 维球面上学习视点、表面法线和三维旋转》，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 9759 - 9767 页，2019 年。5

[69] L. Ladický, B. Zeisl, and M. Pollefeys, ”Discriminatively trained dense surface normal estimation,” in Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6- 12, 2014, Proceedings, Part V 13, pp. 468-484, Springer, 2014. 5, 9

拉迪茨基 L.、蔡斯尔 B. 和波勒菲斯 M.，《判别式训练的密集表面法线估计》，收录于《计算机视觉 - 2014 年欧洲计算机视觉会议 (第 13 届欧洲会议，瑞士苏黎世，2014 年 9 月 6 - 12 日) 论文集，第 V 部分 13》，第 468 - 484 页，施普林格出版社，2014 年。5, 9

[70] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, ”Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1119-1127, 2015. 5

李 B.、沈 C.、戴 Y.、范登亨格尔 A. 和何 M.，《使用深度特征回归和分层条件随机场从单目图像进行深度和表面法线估计》，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 1119 - 1127 页，2015 年。5

[71] X. Long, Y. Zheng, Y. Zheng, B. Tian, C. Lin, L. Liu, H. Zhao, G. Zhou, and W. Wang, ”Adaptive surface normal constraint for geometric estimation from monocular images,” arXiv preprint arXiv:2402.05869, 2024. 5

龙 X.、郑 Y.、郑 Y.、田 B.、林 C.、刘 L.、赵 H.、周 G. 和王 W.，《用于单目图像几何估计的自适应表面法线约束》，预印本 arXiv:2402.05869，2024 年。5

[72] X. Long, C. Lin, L. Liu, W. Li, C. Theobalt, R. Yang, and W. Wang, ”Adaptive surface normal constraint for depth estimation,” in

龙 X.、林 C.、刘 L.、李 W.、特奥博尔特 C.、杨 R. 和王 W.，《用于深度估计的自适应表面法线约束》，收录于

Proceedings of the IEEE/CVF international conference on computer vision, pp. 12849-12858, 2021. 5

《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集》, 第 12849 - 12858 页, 2021 年。5

[73] W. Chen, D. Xiang, and J. Deng, ”Surface normals in the wild,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 1557-1566, 2017. 5

陈 W.、向 D. 和邓 J.,《野外环境下的表面法线》，收录于《电气与电子工程师协会国际计算机视觉会议论文集》，第 1557 - 1566 页，2017 年。5

[74] G. Bae and A. J. Davison, ”Rethinking inductive biases for surface

裴 G. 和戴维森 A. J.,《重新思考表面法线估计的归纳偏置》

normal estimation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2024.5

收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，2024 年。5

[75] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, ”Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8934-8943, 2018. 5

孙 D.、杨 X.、刘 M - Y. 和考茨 J.,《Pwc - net: 使用金字塔、翘曲和代价体积的用于光流的卷积神经网络》，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 8934 - 8943 页，2018 年。5

[76] H. Liu, T. Lu, Y. Xu, J. Liu, and L. Wang, ”Learning optical flow and scene flow with bidirectional camera-lidar fusion,” arXiv preprint arXiv:2303.12017, 2023. 5

刘 H.、陆 T.、徐 Y.、刘 J. 和王 L.,《通过双向相机 - 激光雷达融合学习光流和场景流》，预印本 arXiv:2303.12017，2023 年。5

[77] X. Cheng, P. Wang, and R. Yang, ”Depth estimation via affinity learned with convolutional spatial propagation network,” in Proceedings of the European conference on computer vision (ECCV), pp. 103-119, 2018. 5

程 X.、王 P. 和杨 R.,《通过卷积空间传播网络学习的亲和性进行深度估计》，收录于《欧洲计算机视觉会议 (ECCV) 论文集》，第 103 - 119 页，2018 年。5

[78] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, ”Penet: Towards precise and efficient image guided depth completion,” in 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 13656-13662, IEEE, 2021. 5

胡 M.、王 S.、李 B.、宁 S.、范 L. 和龚 X.，《Penet: 迈向精确高效的图像引导深度补全》，收录于《2021 年电气与电子工程师协会国际机器人与自动化会议 (ICRA) 论文集》，第 13656 - 13662 页，电气与电子工程师协会，2021 年。5

[79] G. Xu, X. Wang, X. Ding, and X. Yang, ”Iterative geometry encoding volume for stereo matching,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21919- 21928, 2023. 5

徐 (Xu)、王 (Wang)、丁 (Ding) 和杨 (Yang)，“用于立体匹配的迭代几何编码体积”，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 21919 - 21928 页，2023 年。5

[80] J. E. Lenssen, C. Osendorfer, and J. Masci, ”Deep iterative surface normal estimation,” in Proceedings of the ieee/cvf conference on computer vision and pattern recognition, pp. 11247-11256, 2020. 5

伦森 (Lenssen)、奥森多费尔 (Osendorfer) 和马斯奇 (Masci)，“深度迭代表面法线估计”，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 11247 - 11256 页，2020 年。5

[81] W. Zhao, S. Liu, Y. Wei, H. Guo, and Y.-J. Liu, ”A confidence-based iterative solver of depths and surface normals for deep multi-view stereo,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6168-6177, 2021. 5, 15

赵 (Zhao)、刘 (Liu)、魏 (Wei)、郭 (Guo) 和刘 (Liu)，“基于置信度的深度和表面法线深度多视图立体迭代求解器”，收录于《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集》，第 6168 - 6177 页，2021 年。5, 15

[82] W. Yin, Y. Liu, C. Shen, A. v. d. Hengel, and B. Sun, ”The devil is in the labels: Semantic segmentation from sentences,” arXiv: Comp. Res. Repository, p. 2202.02002, 2022. 5

尹 (Yin)、刘 (Liu)、沈 (Shen)、亨格尔 (Hengel) 和孙 (Sun)，“魔鬼藏在标签里: 从句子进行语义分割”，预印本服务器: 计算机研究库，第 2202.02002 页，2022 年。5

[83] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., ”Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn., pp. 8748-8763, PMLR, 2021. 5

拉德福德 (Radford)、金 (Kim)、哈拉西 (Hallacy)、拉梅什 (Ramesh)、戈 (Goh)、阿加瓦尔 (Agarwal)、萨斯特里 (Sastry)、阿斯凯尔 (Askell)、米什金 (Mishkin)、克拉克 (Clark) 等，“从自然语言监督中学习可迁移的视觉模型”，收录于《国际机器学习会议论文集》，第 8748 - 8763 页，机器学习研究会议录，2021 年。5

[84] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, ”Mseg: A composite dataset for multi-domain semantic segmentation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 2879-2888, 2020. 5

兰伯特 (Lambert)、刘 (Liu)、森纳 (Sener)、海斯 (Hays) 和科尔图恩 (Koltun)，“Mseg: 用于多领域语义分割的复合数据集”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 2879 - 2888 页，2020 年。5

[85] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., ”Dinov2: Learning robust visual features without supervision,” arXiv preprint arXiv:2304.07193, 2023. 5, 8

奥夸布 (Oquab)、达塞 (Darcet)、穆塔卡尼 (Moutakanni)、沃 (Vo)、萨夫拉涅茨 (Szafraniec)、哈利多夫 (Khalidov)、费尔南德斯 (Fernandez)、哈齐扎 (Haziza)、马萨 (Massa)、埃尔 - 努比 (El - Nouby) 等，“Dinov2: 无监督学习鲁棒视觉特征”，预印本服务器预印本:arXiv:2304.07193，2023 年。5, 8

[86] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ”An image is worth 16x16 words: Transformers for image recognition at scale,” Proc. Int. Conf. Learn. Representations, 2021. 5, 8, 10

多索维茨基 (Dosovitskiy)、拜尔 (Beyer)、科列斯尼科夫 (Kolesnikov)、魏森伯恩 (Weissenborn)、翟 (Zhai)、昂特希纳 (Unterthiner)、德赫加尼 (Dehghani)、明德勒 (Minderer)、海戈尔德 (Heigold)、格利 (Gelly)、乌兹科赖特 (Uszkoreit) 和霍尔兹比 (Houlsby)，“一张图像胜似 16x16 个单词: 大规模图像识别的 Transformer 模型”，《国际学习表征会议论文集》，2021 年。5, 8, 10

[87] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ”High-resolution image synthesis with latent diffusion models,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684-10695, 2022. 5, 10

龙巴赫 (Rombach)、布拉特曼 (Blattmann)、洛伦茨 (Lorenz)、埃瑟 (Esser) 和奥默 (Ommer)，“基于潜在扩散模型的高分辨率图像合成”，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 10684 - 10695 页，2022 年。5, 10

[88] Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gam-baretto, S. Hadap, and J.-F. Lalonde, ”A perceptual measure for deep single image camera calibration,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 2354-2363, 2018. 5

Y. 霍尔德 - 杰弗里 (Y. Hold - Geoffroy)、K. 松卡瓦利 (K. Sunkavalli)、J. 艾森曼 (J. Eisenmann)、M. 费舍尔 (M. Fisher)、E. 甘巴雷托 (E. Gam - baretto)、S. 哈达普 (S. Hadap) 和 J. - F. 拉隆德 (J. - F. Lalonde)，“深度单图像相机校准的感知度量”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 2354 - 2363 页，2018 年。5

[89] M. Lopez-Antequera, P. Gargallo, M. Hofinger, S. R. Bulò, Y. Kuang, and P. Kontschieder, ”Mapillary planet-scale depth dataset,” in Proc. Eur. Conf. Comp. Vis., vol. 12347, pp. 589-604, 2020.7,10

M. 洛佩斯 - 安特克耶拉 (M. Lopez - Antequera)、P. 加尔加洛 (P. Gargallo)、M. 霍芬格 (M. Hofinger)、S. R. 布洛 (S. R. Bulò)、Y. 匡 (Y. Kuang) 和 P. 孔奇德 (P. Kontschieder)，“Mapillary 行星尺度深度数据集”，收录于《欧洲计算机视觉会议论文集》(Proc. Eur. Conf. Comp. Vis.)，第 12347 卷，第 589 - 604 页，2020 年。7,10

[90] D. Singh and B. Singh, ”Investigating the impact of data normalization on classification performance,” Applied Soft Computing, 2019.8

D. 辛格 (D. Singh) 和 B. 辛格 (B. Singh)，“研究数据归一化对分类性能的影响”，《应用软计算》(Applied Soft Computing)，2019 年。8

[91] J. Li, R. Klein, and A. Yao, ”A two-streamed network for estimating fine-scaled depth maps from single rgb images,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 3372-3380, 2017. 9

J. 李 (J. Li)、R. 克莱因 (R. Klein) 和 A. 姚 (A. Yao)，“用于从单张 RGB 图像估计精细尺度深度图的双流网络”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 3372 - 3380 页，2017 年。9

[92] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, ”Deeper depth prediction with fully convolutional residual networks,” in 2016 Fourth international conference on 3D vision (3DV), pp. 239-248, IEEE, 2016. 9

I. 莱纳 (I. Laina)、C. 鲁普雷希特 (C. Rupprecht)、V. 贝拉吉安尼斯 (V. Belagiannis)、F. 通巴里 (F. Tombari) 和 N. 纳瓦布 (N. Navab)，“使用全卷积残差网络进行更深度的深度预测”，收录于 2016 年第四届三维视觉国际会议 (2016 Fourth international conference on 3D vision (3DV))，第 239 - 248 页，电气与电子工程师协会 (IEEE)，2016 年。9

[93] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, ”Depth anything: Unleashing the power of large-scale unlabeled data,” arXiv:2401.10891, 2024. 9

L. 杨 (L. Yang)、B. 康 (B. Kang)、Z. 黄 (Z. Huang)、X. 徐 (X. Xu)、J. 冯 (J. Feng) 和 H. 赵 (H. Zhao)，“深度万物: 释放大规模无标签数据的力量”，预印本 arXiv:2401.10891，2024 年。9

[94] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, ”Learning monocular depth by distilling cross-domain stereo networks,” in Proc. Eur.

X. 郭 (X. Guo)、H. 李 (H. Li)、S. 易 (S. Yi)、J. 任 (J. Ren) 和 X. 王 (X. Wang)，“通过蒸馏跨域立体网络学习单目深度”，收录于《欧洲

Conf. Comp. Vis., pp. 484-500, 2018. 9

计算机视觉会议论文集》(Conf. Comp. Vis.)，第 484 - 500 页，2018 年。9

[95] D. F. Fouhey, A. Gupta, and M. Hebert, ”Unfolding an indoor origami world,” in Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13, pp. 687-702, Springer, 2014. 9

D. F. 福伊 (D. F. Fouhey)、A. 古普塔 (A. Gupta) 和 M. 赫伯特 (M. Hebert)，“展开室内折纸世界”，收录于《计算机视觉 - 2014 年欧洲计算机视觉会议 (Computer Vision - ECCV 2014): 第 13 届欧洲会议，瑞士苏黎世，2014 年 9 月 6 - 12 日，会议论文集，第六部分 13》，第 687 - 702 页，施普林格出版社 (Springer)，2014 年。9

[96] A. Bansal, B. Russell, and A. Gupta, ”Marr revisited: 2d-3d alignment via surface normal prediction,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5965- 5974, 2016. 9

A. 班萨尔 (A. Bansal)、B. 拉塞尔 (B. Russell) 和 A. 古普塔 (A. Gupta)，“重温马尔理论: 通过表面法线预测实现 2D - 3D 对齐”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proceedings of the IEEE conference on computer vision and pattern recognition)，第 5965 - 5974 页，2016 年。9

[97] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille, ”Surge: Surface regularized geometry estimation from a single image,” Advances in Neural Information Processing Systems, vol. 29, 2016.9

P. 王 (P. Wang)、X. 沈 (X. Shen)、B. 拉塞尔 (B. Russell)、S. 科恩 (S. Cohen)、B. 普赖斯 (B. Price) 和 A. L. 尤利尔 (A. L. Yuille)，“Surge: 从单张图像进行表面正则化几何估计”，《神经信息处理系统进展》(Advances in Neural Information Processing Systems)，第 29 卷，2016 年。9

[98] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, ”Pattern-affinitive propagation across depth, surface normal and semantic segmentation,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4106-4115, 2019. 9

Z. 张 (Z. Zhang)、Z. 崔 (Z. Cui)、C. 徐 (C. Xu)、Y. 严 (Y. Yan)、N. 塞贝 (N. Sebe) 和 J. 杨 (J. Yang)，“跨深度、表面法线和语义分割的模式亲和传播”，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》(Proceedings of the IEEE/CVF conference on computer vision and pattern recognition)，第 4106 - 4115 页，2019 年。9

[99] R. Wang, D. Geraghty, K. Matzen, R. Szeliski, and J.-M. Frahm, ”Vplnet: Deep single view normal estimation with vanishing points and lines,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 689-698, 2020. 9

R. 王 (R. Wang)、D. 杰拉蒂 (D. Geraghty)、K. 马曾 (K. Matzen)、R. 斯泽利斯基 (R. Szeliski) 和 J. - M. 弗拉姆 (J. - M. Frahm)，“Vplnet: 利用消失点和线进行深度单视图法线估计”，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》(Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition)，第 689 - 698 页，2020 年。9

[100] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, ”From big to small: Multi-scale local planar guidance for monocular depth estimation,” arXiv: Comp. Res. Repository, p. 1907.10326, 2019. 9

J. H. 李 (J. H. Lee)、M. - K. 韩 (M. - K. Han)、D. W. 高 (D. W. Ko) 和 I. H. 徐 (I. H. Suh)，“从大到小: 用于单目深度估计的多尺度局部平面引导”，预印本 arXiv: 计算机研究库 (Comp. Res. Repository)，第 1907.10326 页，2019 年。9

[101] O. F. Kar, T. Yeo, A. Atanov, and A. Zamir, ”3d common corruptions and data augmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18963- 18974, 2022. 9

O. F. 卡尔 (O. F. Kar)、T. 杨 (T. Yeo)、A. 阿塔诺夫 (A. Atanov) 和 A. 扎米尔 (A. Zamir)，“3D 常见损坏和数据增强”，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》(Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition)，第 18963 - 18974 页，2022 年。9

[102] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, ”A convnet for the 2020s,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 11976-11986, 2022. 8, 10

刘 (Liu)、毛 (Mao)、吴 (C.-Y. Wu)、费希滕霍费尔 (C. Feichtenhofer)、达雷尔 (T. Darrell) 和谢 (S. Xie)，“面向 2020 年代的卷积网络”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 11976 - 11986 页，2022 年。8, 10

[103] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, ”Vision transformers need registers,” arXiv preprint arXiv:2309.16588, 2023. 8

达塞 (T. Darcet)、奥夸布 (M. Oquab)、梅拉勒 (J. Mairal) 和博亚诺夫斯基 (P. Bojanowski)，“视觉 Transformer 需要寄存器”，预印本 arXiv:2309.16588，2023 年。8

[104] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, ”Aggregated residual transformations for deep neural networks,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 1492-1500, 2017. 10

谢 (S. Xie)、吉尔希克 (R. Girshick)、多尔 (P. Dollár)、涂 (Z. Tu) 和何 (K. He)，“深度神经网络的聚合残差变换”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 1492 - 1500 页，2017 年。10

[105] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, ”Scaling vision transformers,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12104-12113, 2022. 10

翟 (X. Zhai)、科列斯尼科夫 (A. Kolesnikov)、霍尔斯比 (N. Houlsby) 和拜尔 (L. Beyer)，“扩展视觉 Transformer”，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 12104 - 12113 页，2022 年。10

[106] A. Geiger, P. Lenz, and R. Urtasun, ”Are we ready for autonomous driving? the kitti vision benchmark suite,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 3354-3361, IEEE, 2012. 9

盖格 (A. Geiger)、伦茨 (P. Lenz) 和乌尔塔松 (R. Urtasun)，“我们准备好迎接自动驾驶了吗？基蒂视觉基准套件”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 3354 - 3361 页，电气与电子工程师协会，2012 年。9

[107] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, ”Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 5828-5839, 2017. 9, 10

戴 (A. Dai)、张 (A. X. Chang)、萨瓦 (M. Savva)、哈尔伯 (M. Halber)、芬克豪泽 (T. Funkhouser) 和尼斯纳 (M. Nießner)，“扫描网络: 室内场景的丰富注释 3D 重建”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 5828 - 5839 页，2017 年。9, 10

[108] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, ”nuscenes: A multimodal dataset for autonomous driving,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 11621-11631, 2020. 9, 10

凯撒 (H. Caesar)、班基蒂 (V. Bankiti)、朗 (A. H. Lang)、沃拉 (S. Vora)、利翁 (V. E. Liong)、徐 (Q. Xu)、克里什南 (A. Krishnan)、潘 (Y. Pan)、巴尔丹 (G. Baldan) 和贝伊博姆 (O. Beijbom)，“nuScenes: 用于自动驾驶的多模态数据集”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 11621 - 11631 页，2020 年。9, 10

[109] T. Koch, L. Liebel, F. Fraundorfer, and M. Korner, ”Evaluation of cnn-based single-image depth estimation methods,” in Eur. Conf. Comput. Vis. Worksh., pp. 0-0, 2018. 9, 10

科赫 (T. Koch)、利贝尔 (L. Liebel)、弗劳恩多尔弗 (F. Fraundorfer) 和科尔纳 (M. Korner)，“基于卷积神经网络的单图像深度估计方法评估”，收录于《欧洲计算机视觉研讨会论文集》，第 0 - 0 页，2018 年。9, 10

[110] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, et al., ”Diode: A dense indoor and outdoor depth dataset,” arXiv: Comp. Res. Repository, p. 1908.00463, 2019. 9, 10

瓦西列维奇 (I. Vasiljevic)、科尔金 (N. Kolkin)、张 (S. Zhang)、罗 (R. Luo)、王 (H. Wang)、戴 (F. Z. Dai)、丹尼尔 (A. F. Daniele)、莫斯塔贾比 (M. Mostajabi)、巴萨特 (S. Basart)、沃尔特 (M. R. Walter) 等，“二极管: 密集的室内外深度数据集”，预印本 arXiv: 计算机研究库，第 1908.00463 页，2019 年。9, 10

[111] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, ”Pixelwise view selection for unstructured multi-view stereo,” in Proc. Eur. Conf. Comp. Vis., 2016. 9

舍恩贝格尔 (J. L. Schönberger)、郑 (E. Zheng)、波勒菲斯 (M. Pollefeys) 和弗拉姆 (J.-M. Frahm)，“非结构化多视图立体视觉的逐像素视图选择”，收录于《欧洲计算机视觉会议论文集》，2016 年。9

[112] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, ”Tanks and temples: Benchmarking large-scale scene reconstruction,” ACM Trans. Graph., vol. 36, no. 4, pp. 1-13, 2017. 9

克纳皮奇 (A. Knapitsch)、朴 (J. Park)、周 (Q.-Y. Zhou) 和科尔图恩 (V. Koltun)，“坦克与寺庙: 大规模场景重建基准测试”，《美国计算机协会图形学汇刊》，第 36 卷，第 4 期，第 1 - 13 页，2017 年。9

[113] S. Li, X. Wu, Y. Cao, and H. Zha, ”Generalizing to the open world: Deep visual odometry with online adaptation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 13184-13193, 2021. 9

李 (S. Li)、吴 (X. Wu)、曹 (Y. Cao) 和查 (H. Zha)，“向开放世界泛化: 具有在线自适应的深度视觉里程计”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 13184 - 13193 页，2021 年。9

[114] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, ”3d packing for self-supervised monocular depth estimation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2020. 10

吉齐利尼 (V. Guizilini)、安布鲁斯 (R. Ambrus)、皮莱 (S. Pillai)、拉文托斯 (A. Raventos) 和盖东 (A. Gaidon)，“用于自监督单目深度估计的 3D 打包”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，2020 年。10

[115] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, ”Level 5 perception dataset 2020.” https://level-5.global/level5/data/, 2019. 10

R. 凯斯滕 (R. Kesten)、M. 乌斯曼 (M. Usman)、J. 休斯顿 (J. Houston)、T. 潘迪亚 (T. Pandya)、K. 纳德哈穆尼 (K. Nadhamuni)、A. 费雷拉 (A. Ferreira)、M. 袁 (M. Yuan)、B. 洛 (B. Low)、A. 贾因 (A. Jain)、P. 翁德鲁斯卡 (P. Ondruska)、S. 奥马里 (S. Omari)、S. 沙阿 (S. Shah)、A. 库尔卡尼 (A. Kulkarni)、A. 卡扎科娃 (A. Kazakova)、C. 陶 (C. Tao)、L. 普拉廷斯基 (L. Platinsky)、W. 江 (W. Jiang) 和 V. 谢特 (V. Shet)，“2020 年 5 级感知数据集”。https://level - 5.global/level5/data/，2019 年 10 月。

[116] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou,

杨 (Yang)、宋 (Song)、黄 (Huang)、邓 (Deng)、施 (Shi) 和周 (Zhou)

”Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2019. 10

“Drivingstereo: 用于自动驾驶场景中立体匹配的大规模数据集”，发表于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，2019 年，第 10 页

[117] J. Cho, D. Min, Y. Kim, and K. Sohn, ”DIML/CVL RGB-D dataset: 2 m RGB-D images of natural indoor and outdoor scenes,” arXiv: Comp. Res. Repository, 2021. 10

赵 (Cho)、闵 (Min)、金 (Kim) 和孙 (Sohn)，“DIML/CVL RGB-D 数据集: 2 m 自然室内外场景的 RGB-D 图像”，预印本服务器 arXiv: 计算机研究库，2021 年，第 10 页。

[118] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, ”Argoverse 2: Next generation datasets for self-driving perception and forecasting,” in Proc. Advances in Neural Inf. Process. Syst., 2021. 10

B. 威尔逊 (B. Wilson)、W. 齐 (W. Qi)、T. 阿加瓦尔 (T. Agarwal)、J. 兰伯特 (J. Lambert)、J. 辛格 (J. Singh)、S. 坎德尔瓦尔 (S. Khandelwal)、B. 潘 (B. Pan)、R. 库马尔 (R. Kumar)、A. 哈特尼特 (A. Hartnett)、J. K. 庞特斯 (J. K. Pontes)、D. 拉马南 (D. Ramanan)、P. 卡尔 (P. Carr) 和 J. 海斯 (J. Hays)，《Argoverse 2: 用于自动驾驶感知和预测的下一代数据集》，收录于《神经信息处理系统进展大会论文集》，2021 年，第 10 页。

[119] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, ”The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016. 10

M. 科茨 (M. Cordts)、M. 奥姆兰 (M. Omran)、S. 拉莫斯 (S. Ramos)、T. 雷菲尔德 (T. Rehfeld)、M. 恩茨韦勒 (M. Enzweiler)、R. 贝嫩森 (R. Benenson)、U. 弗兰克 (U. Franke)、S. 罗斯 (S. Roth) 和 B. 席勒 (B. Schiele),《用于语义城市场景理解的城市景观数据集》，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，2016 年，第 10 页。

[120] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, ”Dsec: A stereo event camera dataset for driving scenarios,” IEEE Robotics and Automation Letters, 2021. 10

M. 格里格 (Gehrig)、W. 阿伦茨 (Aarents)、D. 格里格 (Gehrig) 和 D. 斯卡拉穆扎 (Scaramuzza),“Dsec: 用于驾驶场景的立体事件相机数据集”,《IEEE 机器人与自动化快报》，2021 年，第 10 期。

[121] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, Y. Wang, and D. Yang, ”Pandaset: Advanced sensor suite dataset for autonomous driving,” in IEEE Int. Intelligent Transportation Systems Conf., 2021. 10

肖鹏 (P. Xiao)、邵志 (Z. Shao)、郝帅 (S. Hao)、张泽 (Z. Zhang)、柴鑫 (X. Chai)、焦健 (J. Jiao)、李泽 (Z. Li)、吴杰 (J. Wu)、孙凯 (K. Sun)、蒋凯 (K. Jiang)、王宇 (Y. Wang) 和杨迪 (D. Yang),“Pandaset: 用于自动驾驶的先进传感器套件数据集”，收录于《电气与电子工程师协会国际智能交通系统会议论文集》，2021 年，第 10 页。

[122] Z. Bauer, F. Gomez-Donoso, E. Cruz, S. Orts-Escolano, and M. Cazorla, ”Uasol, a large-scale high-resolution outdoor stereo dataset,” Scientific data, vol. 6, no. 1, pp. 1-14, 2019. 10

Z. 鲍尔 (Z. Bauer)、F. 戈麦斯 - 多诺索 (F. Gomez-Donoso)、E. 克鲁兹 (E. Cruz)、S. 奥尔茨 - 埃斯科拉诺 (S. Orts-Escolano) 和 M. 卡索拉 (M. Cazorla),《Uasol，一个大规模高分辨率户外立体数据集》,《科学数据》(Scientific data)，第 6 卷，第 1 期，第 1 - 14 页，2019 年。10

[123] Y. Cabon, N. Murray, and M. Humenberger, ”Virtual kitti 2,” arXiv preprint arXiv:2001.10773, 2020. 10

Y. 卡邦 (Y. Cabon)、N. 默里 (N. Murray) 和 M. 胡门伯格 (M. Humenberger),《虚拟基蒂数据集 2(Virtual kitti 2)》，预印本 arXiv:2001.10773，2020 年。10

[124] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., ”Scalability in perception for autonomous driving: Waymo open dataset,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2446-2454, 2020. 10

孙 (Sun)、克雷奇马尔 (Kretzschmar)、多蒂瓦拉 (Dotiwalla)、舒阿尔 (Chouard)、帕特奈克 (Patnaik)、崔 (Tsui)、郭 (Guo)、周 (Zhou)、柴 (Chai)、凯恩 (Caine) 等人，“自动驾驶感知的可扩展性:Waymo 开放数据集”，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 2446 - 2454 页，2020 年。10

[125] A. Zamir, A. Sax,, W. Shen, L. Guibas, J. Malik, and S. Savarese, ”Taskonomy: Disentangling task transfer learning,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., IEEE, 2018. 10

A. 扎米尔 (A. Zamir)、A. 萨克斯 (A. Sax)、W. 沈 (W. Shen)、L. 吉巴斯 (L. Guibas)、J. 马利克 (J. Malik) 和 S. 萨瓦雷塞 (S. Savarese),《任务分类学: 解开任务迁移学习的谜团》(“Taskonomy: Disentangling task transfer learning”),收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.),电气与电子工程师协会 (IEEE),2018 年,第 10 页。

[126] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, et al., ”The replica dataset: A digital replica of indoor spaces,” arXiv preprint arXiv:1906.05797, 2019.10

J. 斯特劳布 (J. Straub)、T. 惠兰 (T. Whelan)、L. 马 (L. Ma)、Y. 陈 (Y. Chen)、E. 维伊曼斯 (E. Wijmans)、S. 格林 (S. Green)、J. J. 恩格尔 (J. J. Engel)、R. 穆尔 - 阿塔尔 (R. Mur-Artal)、C. 任 (C. Ren)、S. 维尔马 (S. Verma) 等,《副本数据集: 室内空间的数字副本》,预印本 arXiv:1906.05797,2019 年 10 月

[127] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, et al., ”Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3 d environments for embodied ai,” arXiv preprint arXiv:2109.08238, 2021. 10

S. K. 拉马克里什南 (S. K. Ramakrishnan)、A. 戈卡斯兰 (A. Gokaslan)、E. 维伊曼斯 (E. Wijmans)、O. 马克西梅茨 (O. Maksymets)、A. 克莱格 (A. Clegg)、J. 特纳 (J. Turner)、E. 昂德桑德 (E. Undersander)、W. 加卢巴 (W. Galuba)、A. 韦斯特伯里 (A. Westbury)、A. X. 张 (A. X. Chang) 等人,“栖息地 - Matterport 3D 数据集 (Habitat - Matterport 3D dataset,HM3D): 用于具身人工智能的 1000 个大规模 3 d 环境”,预印本 arXiv:2109.08238,2021 年 10 月。

[128] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, ”Hypersim: A photore-alistic synthetic dataset for holistic indoor scene understanding,” in Proceedings of the IEEE/CVF international conference on computer vision, pp. 10912-10922, 2021. 10

M. 罗伯茨 (M. Roberts)、J. 拉马普拉姆 (J. Ramapuram)、A. 兰詹 (A. Ranjan)、A. 库马尔 (A. Kumar)、M. A. 包蒂斯塔 (M. A. Bautista)、N. 帕赞 (N. Paczan)、R. 韦伯 (R. Webb) 和 J. M. 萨斯金德 (J. M. Susskind),“Hypersim: 用于整体室内场景理解的照片级真实合成数据集”,收录于《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集》,第 10912 - 10922 页,2021 年 10 月。

[129] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, ”A multi-view stereo benchmark with high-resolution images and multi-camera videos,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 3260-3269, 2017. 10

T. 肖普斯 (T. Schops)、J. L. 舍恩贝格尔 (J. L. Schonberger)、S. 加利尼亚尼 (S. Galliani)、T. 萨特勒 (T. Sattler)、K. 辛德勒 (K. Schindler)、M. 波勒菲斯 (M. Pollefeys) 和 A. 盖格 (A. Geiger),“具有高分辨率图像和多相机视频的多视图立体基准测试”,收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》,第 3260 - 3269 页,2017 年 10 月。

[130] J. Kopf, X. Rong, and J.-B. Huang, ”Robust consistent video depth estimation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2021. 12, 13

J. 科普夫 (J. Kopf)、X. 荣 (X. Rong) 和 J. - B. 黄 (J. - B. Huang),“鲁棒一致的视频深度估计”,收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》,2021 年 12 月、13 月。

[131] J.-W. Bian, H. Zhan, N. Wang, T.-J. Chin, C. Shen, and I. Reid, ”Auto-rectify network for unsupervised indoor depth estimation,” IEEE Trans. Pattern Anal. Mach. Intell., 2021. 12, 13

J. - W. 边 (J. - W. Bian)、H. 詹 (H. Zhan)、N. 王 (N. Wang)、T. - J. 钱 (T. - J. Chin)、C. 沈 (C. Shen) 和 I. 里德 (I. Reid)，“用于无监督室内深度估计的自动校正网络”，《电气与电子工程师协会模式分析与机器智能汇刊》，2021 年 12 月、13 月。

[132] M. Sayed, J. Gibson, J. Watson, V. Prisacariu, M. Firman, and C. Godard, ”Simplerecon: 3d reconstruction without 3d convolutions,” in European Conference on Computer Vision, pp. 1-19, Springer, 2022. 12, 13

M. 赛义德 (M. Sayed)、J. 吉布森 (J. Gibson)、J. 沃森 (J. Watson)、V. 普里萨卡里乌 (V. Prisacariu)、M. 菲尔曼 (M. Firman) 和 C. 戈达德 (C. Godard)，“Simplerecon: 无需 3D 卷积的 3D 重建”，收录于《欧洲计算机视觉会议论文集》，第 1 - 19 页，施普林格出版社，2022 年 12 月、13 月。

[133] Z. Yin and J. Shi, ”Geonet: Unsupervised learning of dense depth, optical flow and camera pose,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1983-1992, 2018.13

Z. 尹 (Z. Yin) 和 J. 施 (J. Shi)，“GeoNet: 密集深度、光流和相机位姿的无监督学习”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 1983 - 1992 页，2018 年 13 月。

[134] S. Song, M. Chandraker, and C. C. Guest, ”High accuracy monocular sfm and scale correction for autonomous driving,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 4, pp. 730-743, 2015. 13

S. 宋 (S. Song)、M. 钱德拉克尔 (M. Chandraker) 和 C. C. 格斯特 (C. C. Guest)，“用于自动驾驶的高精度单目运动结构恢复和尺度校正”，《电气与电子工程师协会模式分析与机器智能汇刊》，第 38 卷，第 4 期，第 730 - 743 页，2015 年 13 月。

Supplementary Materials for Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation

Metric3D v2 补充材料: 用于零样本度量深度和表面法线估计的通用单目几何基础模型

January 6, 2025

2025 年 1 月 6 日

1 Details for Models

1 模型详情

Details for ConvNet models. In our work, our encoder employs the ConvNext [1] networks, whose pre-trained weight is from the official released ImageNet- 22k pretraining. The decoder follows the adabins [2]. We set the depth bins number to 256, and the depth range is $[0.3m, 150m]$. We establish 4 flip connections from different levels of encoder blocks to the decoder to merge more low-level features. An hourglass subnetwork is attached to the head of the decoders to enhance background predictions.

卷积神经网络 (ConvNet) 模型详情。在我们的工作中，编码器采用了 ConvNext [1] 网络，其预训练权重来自官方发布的 ImageNet - 22k 预训练。解码器采用 adabins [2]。我们将深度分箱数量设置为 256，深度范围为 $[0.3m, 150m]$ 。我们从编码器块的不同层建立了 4 个跳跃连接到解码器，以融合更多的低层特征。在解码器的头部附加了一个沙漏子网，以增强背景预测。

Details for ViT models. We use dino-v2 transformers [3] with registers [4] as our encoders, which are pre-trained on a curated dataset with 142M images. DPT [5] is used as the decoders. For the ViT-S and ViT-L variants, the DPT decoders take only the last-layer normalized encoder features as the input for stabilized training. The giant ViT-g model instead takes varying-layer features, the same as the original DPT settings. Different from the convnets models above, we use depth bins ranging from $[0.1\text{m}, 200\text{ m}]$ for ViT models.

视觉 Transformer(ViT) 模型详情。我们使用带有寄存器 [4] 的 dino - v2 变换器 [3] 作为编码器，这些编码器在一个包含 142M 张图像的精选数据集上进行了预训练。使用 DPT [5] 作为解码器。对于 ViT - S 和 ViT - L 变体，DPT 解码器仅将最后一层归一化的编码器特征作为输入，以实现稳定训练。而巨型 ViT - g 模型则采用不同层的特征，与原始 DPT 设置相同。与上述卷积神经网络模型不同，我们为 ViT 模型使用的深度分箱范围为 $[0.1\text{m}, 200\text{ m}]$ 。

Details for recurrent blocks. As illustrated in Fig 1. Each recurrent block updates hierarchical features maps $\left\{\mathbf{H}^t_{1/14}, \mathbf{H}^t_{1/7}, \mathbf{H}^t_{1/4}\right\}$ at $\left\{\frac{1}{14}, \frac{1}{7}, \frac{1}{4}\right\}$ scales and the intermediate predictions $\mathbf{S}^t_c, \widehat{\mathbf{N}}^t$ at each iteration step t . This block compromises three Con-vGRU sub-blocks to refine feature maps at different scales, and two projection heads $\mathcal{G}_d$ and $\mathcal{G}_n$ to predict updates for depth and normal respectively. The feature maps are gradually refined from the coarsest $\left(\mathbf{H}^t_{1/14}\right)$ to the finest $\left(\mathbf{H}^t_{1/4}\right)$. For instance, the refined feature map at the $\frac{1}{14}$ scale $\mathbf{H}^{t+1}_{1/14}$ is fed into the second ConvGRU sub-block to refine the $\frac{1}{7}$ -scale feature map $\mathbf{H}^t_{1/7}$. Finally, the projection heads $\mathcal{G}_d, \mathcal{G}_n$ employs a concatenation of original predictions $\widehat{\mathbf{D}}^t_c, \widehat{\mathbf{N}}^t$ and the to finest feature map $\mathbf{H}^t_{1/4}$ to predict the update items $\Delta\widehat{\mathbf{D}}^{t+1}_c, \Delta\widehat{\mathbf{N}}^{t+1}$. Both projection heads are composed of two linear layers with a sandwiched ReLU activation layer.

循环块的详细信息。如图 1 所示。每个循环块在 $\left\{\frac{1}{14}, \frac{1}{7}, \frac{1}{4}\right\}$ 个尺度上更新分层特征图 $\left\{\mathbf{H}^t_{1/14}, \mathbf{H}^t_{1/7}, \mathbf{H}^t_{1/4}\right\}$ ，并在每个迭代步骤 t 更新中间预测结果 $\mathbf{S}^t_c, \widehat{\mathbf{N}}^t$ 。该块包含三个卷积门控循环单元 (ConvGRU) 子块，用于在不同尺度上细化特征图，以及两个投影头 $\mathcal{G}_d$ 和 $\mathcal{G}_n$ ，分别用于预测深度和法线的更新。特征图从最粗糙的 $\left(\mathbf{H}^t_{1/14}\right)$ 逐渐细化到最精细的 $\left(\mathbf{H}^t_{1/4}\right)$ 。例如，$\frac{1}{14}$ 尺度 $\mathbf{H}^{t+1}_{1/14}$ 上的细化特征图被输入到第二个 ConvGRU 子块中，以细化 $\frac{1}{7}$ 尺度的特征图 $\mathbf{H}^t_{1/7}$ 。最后，投影头 $\mathcal{G}_d, \mathcal{G}_n$ 采用原始预测结果 $\widehat{\mathbf{D}}^t_c, \widehat{\mathbf{N}}^t$ 和最精细特征图 $\mathbf{H}^t_{1/4}$ 的拼接来预测更新项 $\Delta\widehat{\mathbf{D}}^{t+1}_c, \Delta\widehat{\mathbf{N}}^{t+1}$ 。两个投影头均由两个线性层和夹在中间的修正线性单元 (ReLU) 激活层组成。

Resource comparison of different models. We compare the resource and performance among our model families in Tab. 1. All inference-time and GPU memory results are computed on an Nvida-A100 40G GPU with the original pytorch implemented models (No engineering optimization like TensorRT or ONNX). Generally, the enormous ViT-Large/giant-backbone models enjoy better performance, while the others are more deployment-friendly. In addition, our models built in classical en-decoder schemes run much faster than the recent diffusion counterpart

[7].

不同模型的资源比较。我们在表 1 中比较了我们模型家族之间的资源和性能。所有推理时间和 GPU 内存结果均在配备原始 PyTorch 实现模型的 Nvidia - A100 40G GPU 上计算 (未进行如 TensorRT 或 ONNX 等工程优化)。一般来说，巨大的视觉 Transformer(ViT) 大/巨型骨干模型具有更好的性能，而其他模型则更便于部署。此外，我们基于经典编码器 - 解码器方案构建的模型比最近的扩散模型 [7] 运行速度快得多。

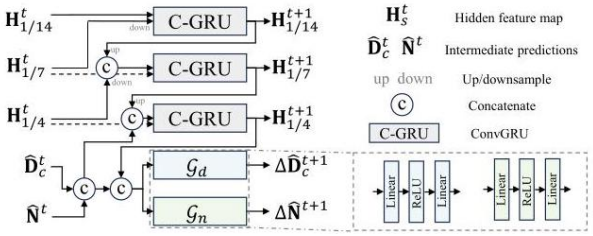

Figure 1: Detailed structure of the update block. Inspired by RAFTStereo [6], we build slow-fast ConvGRU sub-blocks (denoted as 'C-GRU') to refine hierarchical hidden feature maps. Projection heads $\mathcal{G}_d, \mathcal{G}_n$ are attached to the end of the final Con-vGRU to prediction update items for the predictions.

图 1: 更新块的详细结构。受 RAFTStereo [6] 的启发，我们构建了快慢卷积门控循环单元 (ConvGRU) 子块 (表示为“C - GRU”) 来细化分层隐藏特征图。投影头 $\mathcal{G}_d, \mathcal{G}_n$ 连接到最终卷积门控循环单元 (Con - vGRU) 的末尾，用于预测预测结果的更新项。

1.1 Datasets and Training and Testing

1.1 数据集以及训练与测试

We collect over 16M data from 18 public datasets for training. Datasets are listed in Tab. 2. When training the ConvNeXt-backbone models, we use a smaller collection containing the following 11 datasets with 8M images: DDAD [8], Lyft [9], DrivingStereo [10], DIML [11], Argov-erse2 [12], Cityscapes [13], DSEC [14], Maplillary PSD [15], Pandaset [16], UASOL[17], and Taskon-omy [20]. In the autonomous driving datasets, including DDAD [8], Lyft [9], DrivingStereo [10], Ar-goverse2 [12], DSEC [14], and Pandaset [16], have provided LiDar and camera intrinsic and extrinsic parameters. We project the LiDar to image planes to obtain ground-truth depths. In contrast, Cityscapes [13], DIML [11], and UASOL [17] only provide calibrated stereo images. We use raft-stereo [6] to achieve pseudo ground-truth depths. Mapillary PSD [15] dataset provides paired RGB-D, but the depth maps are achieved from a structure-from-motion method. The camera intrinsic parameters are estimated from the SfM. We believe that such achieved metric information is noisy. Thus we do not enforce learning-metric-depth loss on this data, i.e., L_{silog} , to reduce the effect of noises. For the Taskon-omy [20] dataset, we follow LeReS [31] to obtain the instance planes, which are employed in the pair-wise normal regression loss. During training, we employ the training strategy from [32] to balance all datasets in each training batch.

我们从 18 个公开数据集中收集了超过 16M 的数据用于训练。数据集列于表 2。在训练 ConvNeXt 骨干模型时，我们使用一个较小的数据集集合，其中包含以下 11 个数据集，共有 8M 张图像:DDAD [8]、Lyft [9]、DrivingStereo [10]、DIML [11]、Argoverse2 [12]、Cityscapes [13]、DSEC [14]、Maplillary PSD [15]、Pandaset [16]、UASOL [17] 和 Taskonomy [20]。在自动驾驶数据集中，包括 DDAD [8]、Lyft [9]、DrivingStereo [10]、Argoverse2 [12]、DSEC [14] 和 Pandaset [16]，都提供了激光雷达 (LiDar) 以及相机的内参和外参。我们将激光雷达数据投影到图像平面以获取真实深度值。相比之下，Cityscapes [13]、DIML [11] 和 UASOL [17] 仅提供经过校准的立体图像。我们使用 raft-stereo [6] 来获取伪真实深度值。Mapillary PSD [15] 数据集提供了配对的 RGB-D 数据，但深度图是通过运动恢复结构 (structure-from-motion) 方法获得的。相机内参是从运动恢复结构方法中估计得到的。我们认为这样获得的度量信息存在噪声。因此，我们不在此数据上强制使用学习度量深度损失，即 L_{silog}，以减少噪声的影响。对于 Taskonomy [20] 数据集，我们遵循 LeReS [31] 的方法来获取实例平面，这些实例平面用于成对法线回归损失。在训练过程中，我们采用文献 [32] 中的训练策略来平衡每个训练批次中的所有数据集。

The testing data is listed in Tab. 2. All of them are captured by high-quality sensors. In testing, we employ their provided camera intrinsic parameters to perform our proposed canonical space transformation.

测试数据列于表 2。所有数据均由高质量传感器采集。在测试时，我们使用它们提供的相机内参来执行我们提出的规范空间变换。

1.2 Details for Some Experiments

1.2 部分实验细节

Finetuning protocols. To finetune the large-scale-data trained models on some specific datasets, we use the ADAM optimizer with the initial learning rate beginning at 10^{-6} and linear decayed to 10^{-7} within 6 K steps. Notably, such finetune does not require a large batch size like large-scale training. We use in practice a batch-size of 16 for the ViT-g model and 32 for ViT-L. The models will converge quickly in approximately 2 K steps. To stabilize finetuning, we also leverage the predictions $\widetilde{\mathbf{D}}_c\widetilde{\mathbf{N}}$ of the pre-finetuned model as alternative pseudo labels. These labels can sufficiently impose supervision upon the annotation-absent regions. The complete loss can be formulated as:

微调协议。为了在某些特定数据集上微调大规模数据训练的模型，我们使用 ADAM 优化器，初始学习率从 10^{-6} 开始，并在 6 K 步内线性衰减到 10^{-7}。值得注意的是，这种微调不需要像大规模训练那样大的批量大小。实际上，我们对 ViT - g 模型使用 16 的批量大小，对 ViT - L 使用 32 的批量大小。模型将在大约 2 K 步内快速收敛。为了稳定微调过程，我们还利用预微调模型的预测结果 $\widetilde{\mathbf{D}}_c\widetilde{\mathbf{N}}$ 作为替代伪标签。这些标签可以对无标注区域进行充分的监督。完整的损失函数可以表示为:

$$L_{ft} = 0.01L_{d-n}\left(\mathbf{D}_c, \mathbf{N}\right) + L_d\left(\mathbf{D}_c, \mathbf{D}_c^*\right) + L_n\left(\mathbf{N}, \mathbf{N}^*\right)$$

$$+0.01\left(L_d\left(\mathbf{D}_c, \widetilde{\mathbf{D}}_c\right) + L_n\left(\mathbf{N}, \widetilde{\mathbf{N}}\right)\right), \tag{1}$$

where $\mathbf{D}_c$ and $\mathbf{N}$ are the predicted depth in the canonical space and surface normal, $\mathbf{D}_c^*$ and $\mathbf{N}^*$ are the groundtruth labels, L_d, L_n , and L_{d-n} are the losses for depth, normal, and depth-normal consistency introduced in the main text.

其中 $\mathbf{D}_c$ 和 $\mathbf{N}$ 分别是规范空间中预测的深度和表面法线，$\mathbf{D}_c^*$ 和 $\mathbf{N}^*$ 是真实标签，L_d, L_n 和 L_{d-n} 分别是正文中引入的深度、法线和深度 - 法线一致性损失。

Evaluation of zero-shot 3D scene reconstruction. In this experiment, we use all methods' released models to predict each frame's depth and use the ground-truth poses and camera intrinsic parameters to reconstruct point clouds. When evaluating the reconstructed point cloud, we employ the iterative closest point (ICP) [33] algorithm to match the predicted point clouds with ground truth by a pose transformation matrix. Finally, we evaluate the Chamfer ℓ_1 distance and F-score on the point cloud. Reconstruction of in-the-wild scenes. We collect several photos from Flickr. From their associated camera metadata, we can obtain the focal length $\widehat{f}$ and the pixel size δ . According to $\widehat{f}/\delta$, we can obtain the pixel-represented focal length for 3D reconstruction and achieve the metric information. We use meshlab software to measure some structures' size on point clouds. More visual results are shown in Fig. 7.

零样本 3D 场景重建评估。在这个实验中，我们使用所有方法发布的模型来预测每一帧的深度，并使用真实姿态和相机内参来重建点云。在评估重建的点云时，我们使用迭代最近点 (ICP)[33] 算法，通过一个姿态变换矩阵将预测的点云与真实点云进行匹配。最后，我们在点云上评估倒角 ℓ_1 距离和 F 分数。野外场景重建。我们从 Flickr 上收集了几张照片。从它们相关的相机元数据中，我们可以获得焦距 $\widehat{f}$ 和像素尺寸 δ 。根据 $\widehat{f}/\delta$ ，我们可以获得用于 3D 重建的以像素表示的焦距，并获得度量信息。我们使用 meshlab 软件来测量点云上某些结构的尺寸。更多可视化结果如图 7 所示。

Generalization of metric depth estimation. To evaluate our method's robustness of metric recovery, we test on 7 zero-shot datasets, i.e. NYU, KITTI, DIODE (indoor and outdoor parts), ETH3D, iBims- 1, and NuScenes. Details are reported in Tab. 2. We use the officially provided focal length to predict the metric depths. All benchmarks use the same depth model for evaluation. We don't perform any scale alignment.

度量深度估计的泛化。为了评估我们方法在度量恢复方面的鲁棒性，我们在 7 个零样本数据集上进行了测试，即纽约大学数据集 (NYU)、基蒂数据集 (KITTI)、二极管数据集 (DIODE，包括室内和室外部分)、苏黎世联邦理工学院 3D 数据集 (ETH3D)、iBims - 1 数据集和努场景数据集 (NuScenes)。具体细节见表 2。我们使用官方提供的焦距来预测度量深度。所有基准测试都使用相同的深度模型进行评估。我们不进行任何尺度对齐。

Table 1: Comparative analysis of resource and performance across our model families includes evaluation of resource utilization metrics such as inference speed, memory usage, and the proportion of optimization modules. Additionally, we assess metric depth performance on KITTI/NYU datasets and normal performance on NYUv2 dataset, with results derived from checkpoints without fine-tuning. For ViT models, inference speed is measured using 16-bit precision (Bfloat16), which is the same precision as the training setup.

表 1: 对我们模型族的资源和性能进行比较分析，包括对推理速度、内存使用和优化模块比例等资源利用指标的评估。此外，我们评估了在基蒂/纽约大学数据集 (KITTI/NYU) 上的度量深度性能以及在纽约大学 v2 数据集 (NYUv2) 上的法线性能，结果来自未经微调的检查点。对于视觉 Transformer(ViT) 模型，推理速度使用 16 位精度 (Bfloat16) 进行测量，这与训练设置的精度相同。

Model			Resource			KITTI Depth		NYUv2 Depth		NYUv2 Normal	
Encoder	Decoder	Optim.	Speed	GPU Memory	Optim. time	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	Median↓ 30° ↑	
Marigold[7] VAE+U-net	U-net+VAE	-	0.13 fps	17.3G	-	No metric	No metric	No metric	No metric	-	-
Ours ConvNeXt-Large	Hourglass	-	10.5 fps	4.2G	-	0.053	0.965	0.083	0.944	-	-
Ours ViT-Small	DPT	4 steps	11.6 fps	2.9G	3.4%	0.070	0.937	0.084	0.945	7.7	0.870
Ours ViT-Large	DPT	8 steps	9.5 fps	7.0G	9.5%	0.052	0.974	0.063	0.975	7.0	0.881
Ours ViT-giant	DPT	8 steps	5.0 fps	15.6G	25%	0.051	0.977	0.067	0.980	7.1	0.881

模型			资源			KITTI 深度数据集		NYUv2 深度数据集		NYUv2 法线数据集	
编码器	解码器	优化器	速度	GPU 内存	优化时间	绝对相对误差 ↓	δ_1 ↑	绝对相对误差 ↓	δ_1 ↑	中位数 ↓ 30° ↑	
金盏花 [7] 变分自编码器 +U 型网络	U 型网络 + 变分自编码器	-	0.13 帧/秒	17.3G	-	无指标	无指标	无指标	无指标	-	-
我们的 ConvNeXt-Large 模型	沙漏网络	-	10.5 帧/秒	4.2G	-	0.053	0.965	0.083	0.944	-	-
我们的 ViT-Small 模型	DPT 模型	4 步	11.6 帧/秒	2.9G	3.4%	0.070	0.937	0.084	0.945	7.7	0.870
我们的 ViT-Large 模型	DPT 模型	8 步	9.5 帧/秒	7.0G	9.5%	0.052	0.974	0.063	0.975	7.0	0.881
我们的 ViT-giant 模型	DPT 模型	8 步	5.0 帧/秒	15.6G	25%	0.051	0.977	0.067	0.980	7.1	0.881

Evaluation on affine-invariant depth benchmarks. We follow existing affine-invariant depth estimation methods to evaluate 5 zero-shot datasets. Before evaluation, we employ the least square fitting to align the scale and shift with ground truth [34]. Previous methods' performance is cited from their papers.

仿射不变深度基准评估。我们遵循现有的仿射不变深度估计方法对 5 个零样本数据集进行评估。在评估之前，我们采用最小二乘法拟合来使尺度和偏移与真实值对齐 [34]。先前方法的性能引用自它们的论文。

Dense-SLAM Mapping. This experiment is conducted on the KITTI odometry benchmark. We use Table 2: Training and testing datasets used for experiments. our model to predict metric depths, and then naively input them to the Droid-SLAM system as an initial depth. We do not perform any finetuning but directly run their released codes on KITTI. With Droid-SLAM predicted poses, we unproject depths to the 3D point clouds and fuse them together to achieve dense metric mapping. More qualitative results are shown in Fig. 6.

稠密同步定位与地图构建 (Dense - SLAM) 建图。本实验在 KITTI 里程计基准数据集上进行。我们使用表 2: 实验所用的训练和测试数据集。我们的模型来预测度量深度，然后简单地将其作为初始深度输入到 Droid - SLAM 系统中。我们不进行任何微调，而是直接在 KITTI 上运行他们发布的代码。利用 Droid - SLAM 预测的位姿，我们将深度反投影到 3D 点云并将它们融合在一起以实现稠密度量建图。更多定性结果如图 6 所示。

Datasets	Scenes	Source	Label	Size	#Cam.
Training Data					
DDAD [8]	Outdoor	Real-world	Depth	~ 80 K	36+
Lyft [9]	Outdoor	Real-world	Depth	~ 50 K	6+
Driving Stereo (DS) [10]	Outdoor	Real-world	Depth	181K	1
DIML [11]	Outdoor	Real-world	Depth	~ 122K	10
Arogoverse2 [12]	Outdoor	Real-world	Depth	3515K	6+
Cityscapes [13]	Outdoor	Real-world	Depth	170K	1
DSEC [14]	Outdoor	Real-world	Depth	26K	1
Mapillary PSD [15]	Outdoor	Real-world	Depth	750K	1000+
Pandaset [16]	Outdoor	Real-world	Depth	48K	6
UASOL [17]	Outdoor	Real-world	Depth	1370K	1
Virtual KITTI [18]	Outdoor	Synthesized	Depth	37K	2
Waymo [19]	Outdoor	Real-world	Depth	1M	5
Matterport3d [20]	In/Out	Real-world	Depth + Normal	144K	3
Taskonomy [20]	Indoor	Real-world	Depth + Normal	4M	$\sim$ 1M
Replica [21]	Indoor	Real-world	Depth + Normal	~ 150K	1
ScanNet † [22]	Indoor	Real-world	Depth + Normal	2.5M	1
HM3d [23]	Indoor	Real-world	Depth + Normal	2000K	1
Hypersim [24]	Indoor	Synthesized	Depth + Normal	54K	1
Testing Data					
NYU [25]	Indoor	Real-world	Depth+Normal	654	1
KITTI [26]	Outdoor	Real-world	Depth	652	4
ScanNet † [22]	Indoor	Real-world	Depth+Normal	700	1
NuScenes (NS) [27]	Outdoor	Real-world	Depth	10K	6
ETH3D [28]	Outdoor	Real-world	Depth	431	1
DIODE [29]	In/Out	Real-world	Depth	771	1
iBims-1 [30]	Indoor	Real-world	Depth	100	1

数据集	场景	来源	标签	大小	摄像头数量
训练数据					
DDAD [8]	室外	真实世界	深度	∼ 80 K	36+
Lyft [9]	室外	真实世界	深度	∼ 50 K	6+
驾驶立体数据集 (Driving Stereo，DS) [10]	室外	真实世界	深度	ĩ81K	1
DIML [11]	室外	真实世界	深度	∼ 122K	10
Arogoverse2 [12]	室外	真实世界	深度	ẽ515K	6+
城市景观数据集 (Cityscapes) [13]	室外	真实世界	深度	ĩ70K	1
DSEC [14]	室外	真实世界	深度	ẽ6K	1
Mapillary PSD [15]	室外	真实世界	深度	750K	1000+
Pandaset [16]	室外	真实世界	深度	ã8K	6
UASOL [17]	室外	真实世界	深度	ĩ370K	1
虚拟 KITTI 数据集 (Virtual KITTI) [18]	室外	合成的	深度	37K	2
Waymo [19]	室外	真实世界	深度	ĩM	5
Matterport3d [20]	室内/室外	真实世界	深度 + 法线	144K	3
任务分类数据集 (Taskonomy) [20]	室内	真实世界	深度 + 法线	ã4M	∼ 1M
Replica [21]	室内	真实世界	深度 + 法线	∼ 150K	1
扫描网络数据集 (ScanNet) † [22]	室内	真实世界	深度 + 法线	ẽ2.5M	1
HM3d [23]	室内	真实世界	深度 + 法线	ẽ000K	1
Hypersim [24]	室内	合成的	深度 + 法线	54K	1
测试数据					
纽约大学数据集 (NYU) [25]	室内	真实世界	深度 + 法线	654	1
基蒂视觉基准数据集 (KITTI) [26]	室外	真实世界	深度	652	4
扫描网络数据集 (ScanNet) † [22]	室内	真实世界	深度 + 法线	700	1
NuScenes(NS) [27]	室外	真实世界	深度	10K	6
ETH3D [28]	室外	真实世界	深度	431	1
DIODE [29]	室内/室外	真实世界	深度	771	1
iBims - 1 [30]	室内	真实世界	深度	100	1

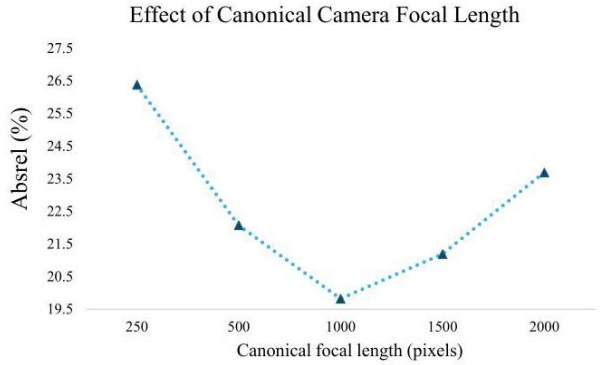

Figure 2: Effect of varying canonical focal lengths. We apply different canonical focal lengths and find that an intermediate focal length leads to the best performance.

图 2: 不同标准焦距的影响。我们应用不同的标准焦距，发现中等焦距能带来最佳性能。

Ablation on canonical space. To study the effect of different focal lengths in the canonical space, we train the ConvNext models on the small subset of varying datasets and test their performance on the validation set. We compare the absrel error using $\{250, 500, 1000, 1500$, and 2000 $\}$ -pixel canonical camera focal lengths. As

shown in Fig. 2, the canonical focal length of 1000 achieves the lowest depth er- † ScanNet is a non-zero-shot testing dataset for our ViT models. ror. We apply this setting for all other experiments.

标准空间的消融实验。为了研究标准空间中不同焦距的影响，我们在不同数据集的小子集上训练 ConvNext 模型，并在验证集上测试其性能。我们比较了使用 $\{250, 500, 1000, 1500$ 以及 2000 $\}$ 像素标准相机焦距时的绝对相对误差 (absrel error)。如图 2 所示，1000 的标准焦距实现了最低的深度误差。† ScanNet 是我们 ViT 模型的非零样本测试数据集。我们在所有其他实验中都采用了这一设置。

1.3 More Visual Results

1.3 更多可视化结果

Qualitative comparison of depth and normal estimation. In Figs 3, 4, we compare visualized depth and normal maps from the Vit-g CSTM_label model with ZoeDepth [35], Bae etal [36], and Om-nidata [37]. In Figs. 5, 10, 11, and 12, We show the qualitative comparison of our depth maps from the ConvNeXt-L CSTM_label model with Adabins [2], NewCRFs [38], and Omnidata [37]. Our results have much fine-grained details and less artifacts.

深度和法线估计的定性比较。在图 3、图 4 中，我们将 Vit - g CSTM_label 模型的可视化深度图和法线图与 ZoeDepth [35]、Bae 等人 [36] 以及 Omnidata [37] 进行了比较。在图 5、图 10、图 11 和图 12 中，我们展示了 ConvNeXt - L CSTM_label 模型的深度图与 Adabins [2]、NewCRFs [38] 以及 Omnidata [37] 的定性比较。我们的结果具有更精细的细节和更少的伪影。

Visualization of iterative refinement. To comprehensively understand the usage of iterative refinement modules, we visualize the predictions before/after optimization and the updates for different steps in Fig. 8. Here we use our publicly released 4- step ViT-small model for visualization. Initially, the network produces a coarse prediction. The first step updates the most drastically, while sub-sequential steps focus mainly on object boundaries. Finally, the refined predictions have clearer shapes and sharper edges.

迭代细化的可视化。为了全面理解迭代细化模块的使用，我们在图 8 中可视化了优化前后的预测结果以及不同步骤的更新情况。这里我们使用公开发布的 4 步 ViT - small 模型进行可视化。最初，网络产生一个粗略的预测。第一步的更新最为剧烈，而后续步骤主要关注物体边界。最后，细化后的预测结果具有更清晰的形状和更锐利的边缘。

Reconstructing 360° NuScenes scenes. Current autonomous driving cars are equipped with several pin-hole cameras to capture 360° views. Capturing the surround-view depth is important for autonomous driving. We sampled some scenes from the testing data of NuScenes. With our depth model, we can obtain the metric depths for 6 -ring cameras. With the provided camera intrinsic and extrinsic parameters, we unproject the depths to the 3D point cloud and merge all views together. See Fig. 9 for details. Note that 6-ring cameras have different camera intrinsic parameters. We can observe that all views' point clouds can be fused together consistently.

重建 360° NuScenes 场景。当前的自动驾驶汽车配备了多个针孔相机来捕捉 360° 视角。捕捉环视深度信息对自动驾驶至关重要。我们从 NuScenes 的测试数据中采样了一些场景。使用我们的深度模型，我们可以获得 6 环相机的度量深度。利用提供的相机内参和外参，我们将深度信息反投影到 3D 点云并将所有视角合并在一起。详情见图 9。请注意，6 环相机具有不同的相机内参。我们可以观察到所有视角的点云能够一致地融合在一起。

References

参考文献

[1] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, ”A convnet for the 2020s,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 11976- 11986, 2022.

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “21 世纪 20 年代的卷积网络 (A convnet for the 2020s)”，见《电气与电子工程师协会计算机视觉与模式识别会议论文集 (Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)》，第 11976 - 11986 页，2022 年。

[2] S. F. Bhat, I. Alhashim, and P. Wonka, ”Adabins: Depth estimation using adaptive bins,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 4009- 4018, 2021.

S. F. Bhat, I. Alhashim, and P. Wonka,“自适应分箱的深度估计 (Adabins: Depth estimation using adaptive bins)”，见《电气与电子工程师协会计算机视觉与模式识别会议论文集 (Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)》，第 4009 - 4018 页，2021 年。

[3] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., ”Dinov2: Learning robust visual features without supervision,” arXiv preprint arXiv:2304.07193, 2023.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El - Nouby 等，“Dinov2: 无监督学习鲁棒视觉特征 (Dinov2: Learning robust visual features without supervision)”，预印本 arXiv:2304.07193，2023 年。

[4] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, ”Vision transformers need registers,” arXiv preprint arXiv:2309.16588, 2023.

T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, “视觉 Transformer 需要寄存器 (Vision transformers need registers)”，预印本 arXiv:2309.16588，2023 年。

[5] R. Ranftl, A. Bochkovskiy, and V. Koltun, ”Vision transformers for dense prediction,” in Proc. IEEE Int. Conf. Comp. Vis., pp. 12179-12188, 2021.

R. Ranftl, A. Bochkovskiy, and V. Koltun, “用于密集预测的视觉 Transformer(Vision transformers for dense prediction)”，见《电气与电子工程师协会国际计算机视觉会议论文集 (Proc. IEEE Int. Conf. Comp. Vis.)》，第 12179 - 12188 页，2021 年。

[6] L. Lipson, Z. Teed, and J. Deng, ”Raft-stereo: Multilevel recurrent field transforms for stereo matching,” in Int. Conf. 3D. Vis., 2021.

L. Lipson, Z. Teed, and J. Deng, “RAFT - 立体匹配: 用于立体匹配的多级循环场变换 (Raft - stereo: Multilevel recurrent field transforms for stereo matching)”, 见《国际三维视觉会议 (Int. Conf. 3D. Vis.)》, 2021 年。

[7] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, ”Repurposing diffusion-based image generators for monocular depth estimation,” arXiv preprint arXiv:2312.02145, 2023.

B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, “将基于扩散的图像生成器用于单目深度估计 (Repurposing diffusion - based image generators for monocular depth estimation)”, 预印本 arXiv:2312.02145, 2023 年。

[8] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, ”3d packing for self-supervised monocular depth estimation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2020.

V. 吉齐利尼 (V. Guizilini)、R. 安布鲁斯 (R. Ambrus)、S. 皮莱 (S. Pillai)、A. 拉文托斯 (A. Raventos) 和 A. 盖登 (A. Gaidon), “用于自监督单目深度估计的 3D 打包方法”, 收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.), 2020 年。

[9] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, ”Level 5 perception dataset 2020.” https://level-5.global/level5/data/, 2019.

R. 凯斯滕 (R. Kesten)、M. 乌斯曼 (M. Usman)、J. 休斯顿 (J. Houston)、T. 潘迪亚 (T. Pandya)、K. 纳德哈穆尼 (K. Nadhamuni)、A. 费雷拉 (A. Ferreira)、M. 袁 (M. Yuan)、B. 洛 (B. Low)、A. 贾因 (A. Jain)、P. 翁德鲁斯卡 (P. Ondruska)、S. 奥马里 (S. Omari)、S. 沙阿 (S. Shah)、A. 库尔卡尼 (A. Kulkarni)、A. 卡扎科娃 (A. Kazakova)、C. 陶 (C. Tao)、L. 普拉廷斯基 (L. Platinsky)、W. 江 (W. Jiang) 和 V. 谢特 (V. Shet), “2020 年 5 级感知数据集”。https://level - 5.global/level5/data/, 2019 年。

[10] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, ”Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2019.

G. 杨 (G. Yang)、X. 宋 (X. Song)、C. 黄 (C. Huang)、Z. 邓 (Z. Deng)、J. 施 (J. Shi) 和 B. 周 (B. Zhou),“Drivingstereo: 用于自动驾驶场景中立体匹配的大规模数据集”, 收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.), 2019 年。

[11] J. Cho, D. Min, Y. Kim, and K. Sohn, ”DIML/CVL RGB-D dataset: 2 m RGB-D images of natural indoor and outdoor scenes,” arXiv: Comp. Res. Repository, 2021.

J. 赵 (J. Cho)、D. 闵 (D. Min)、Y. 金 (Y. Kim) 和 K. 孙 (K. Sohn), “DIML/CVL RGB - D 数据集: 2 m 自然室内外场景的 RGB - D 图像”, 预印本服务器 (arXiv): 计算机研究库, 2021 年。

[12] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, ”Ar-goverse 2: Next generation datasets for self-driving perception and

forecasting,” in Proc. Advances in Neural Inf. Process. Syst., 2021.

B. 威尔逊 (B. Wilson)、W. 齐 (W. Qi)、T. 阿加瓦尔 (T. Agarwal)、J. 兰伯特 (J. Lambert)、J. 辛格 (J. Singh)、S. 坎德尔瓦尔 (S. Khandelwal)、B. 潘 (B. Pan)、R. 库马尔 (R. Kumar)、A. 哈特尼特 (A. Hartnett)、J. K. 庞特斯 (J. K. Pontes)、D. 拉马南 (D. Ramanan)、P. 卡尔 (P. Carr) 和 J. 海斯 (J. Hays)，“Ar - goverse 2: 用于自动驾驶感知和预测的下一代数据集”，收录于《神经信息处理系统进展会议论文集》(Proc. Advances in Neural Inf. Process. Syst.)，2021 年。

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. En-zweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, ”The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

M. 科尔特斯 (M. Cordts)、M. 奥姆兰 (M. Omran)、S. 拉莫斯 (S. Ramos)、T. 雷菲尔德 (T. Rehfeld)、M. 恩茨韦勒 (M. En - zweiler)、R. 贝嫩森 (R. Benenson)、U. 弗兰克 (U. Franke)、S. 罗斯 (S. Roth) 和 B. 席勒 (B. Schiele)，“用于语义城市场景理解的城市景观数据集”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，2016 年。

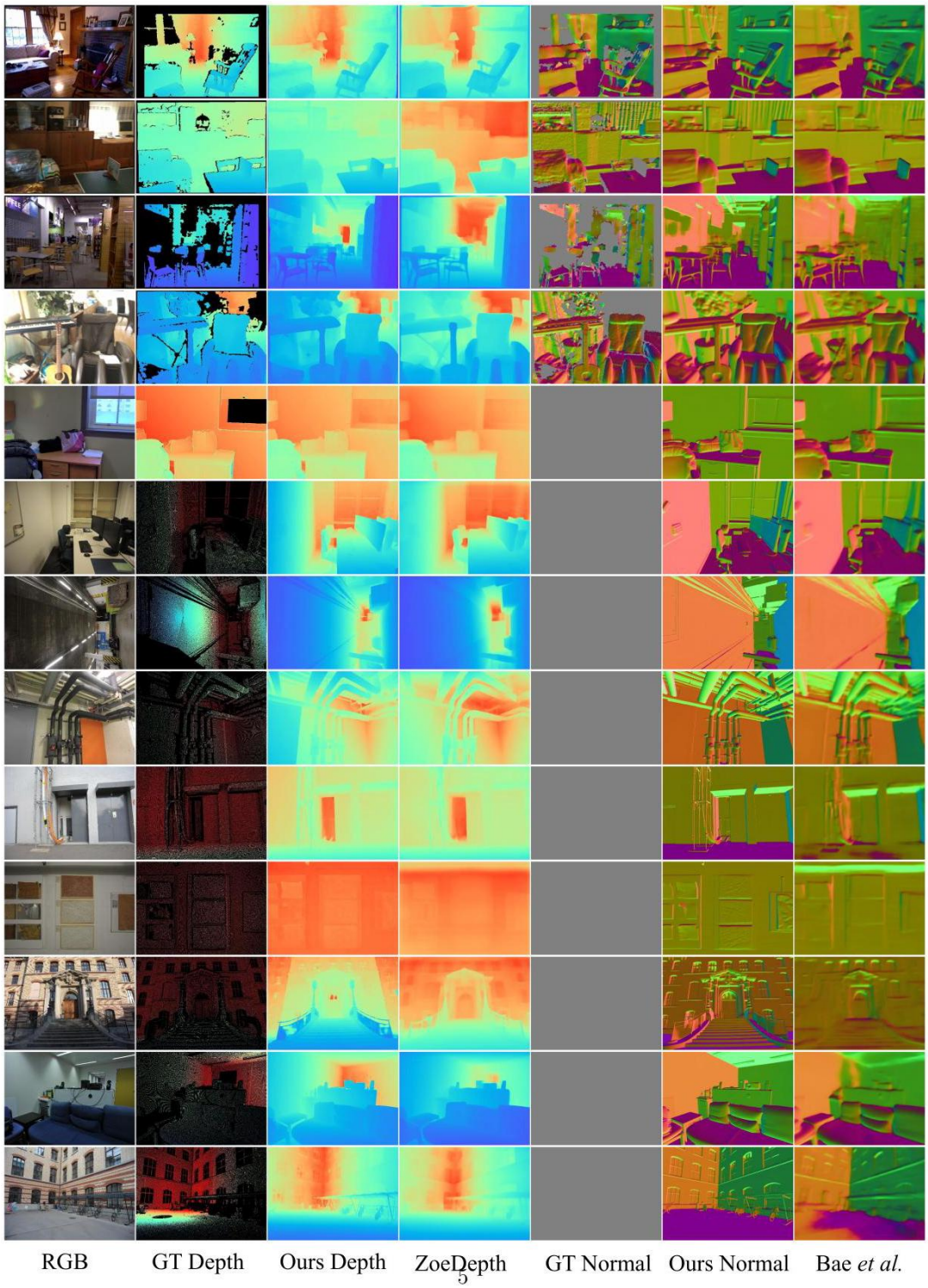

Figure 3: Depth and normal estimation. The visual comparison of predicted depth and normal on indoor/outdoor scenes from NYUv2, iBims, Eth3d, and ScanNet. Our depth and normal maps come from the ViT-g CSTM_label model.

图 3: 深度和法线估计。对来自 NYUv2、iBims、Eth3d 和 ScanNet 的室内/室外场景的预测深度和法线进行可视化比较。我们的深度图和法线图来自 ViT - g CSTM_label 模型。

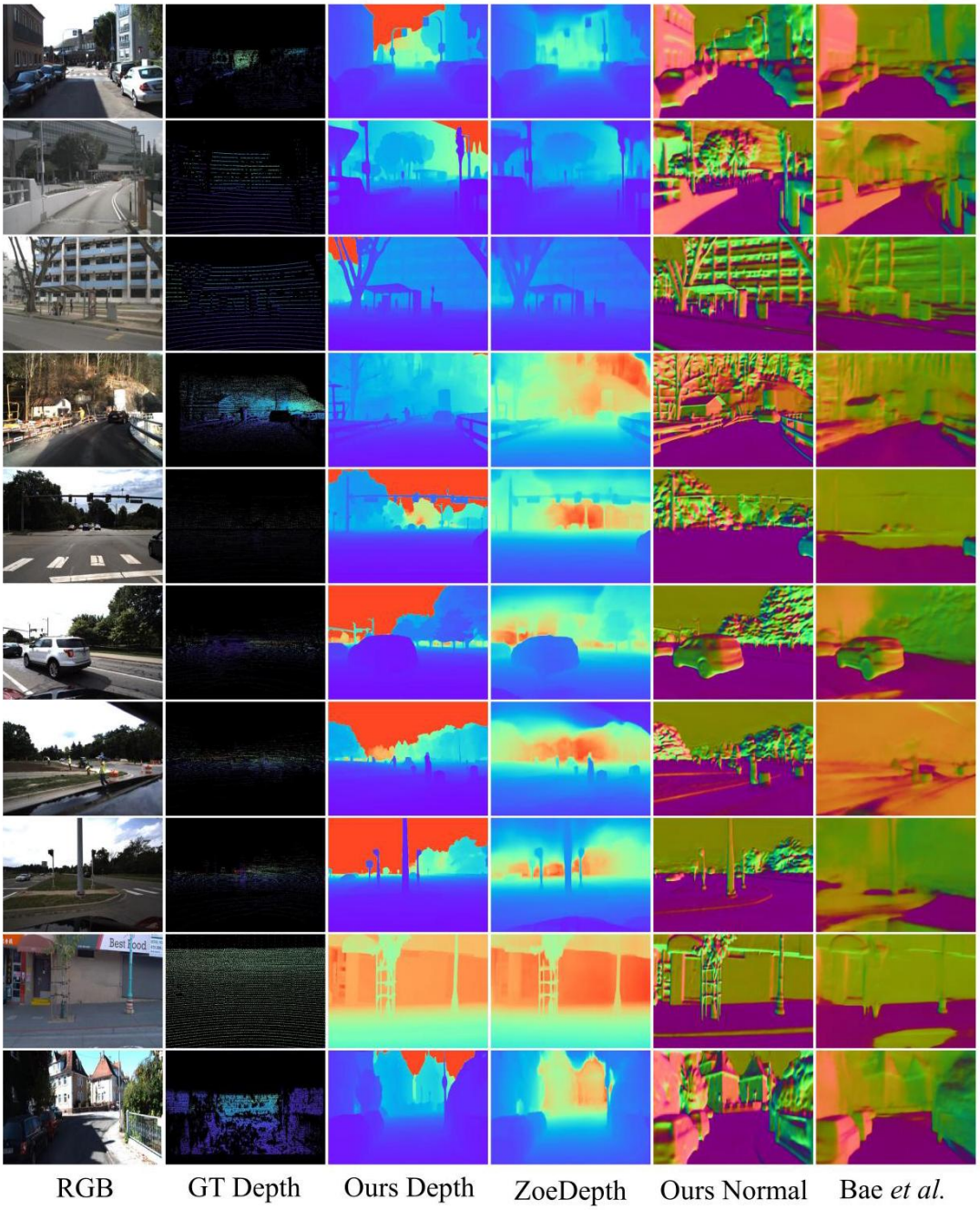

Figure 4: Depth and normal estimation. The visual comparison of predicted depth and normal on driving scenes from KITTI, Nuscenes, DIML, DDAD, and Waymo. Our depth and normal maps come from the ViT-g CSTM_label model.

图 4: 深度和法线估计。对来自 KITTI、Nuscenes、DIML、DDAD 和 Waymo 的驾驶场景的预测深度和法线进行可视化比较。我们的深度图和法线图来自 ViT - g CSTM_label 模型。

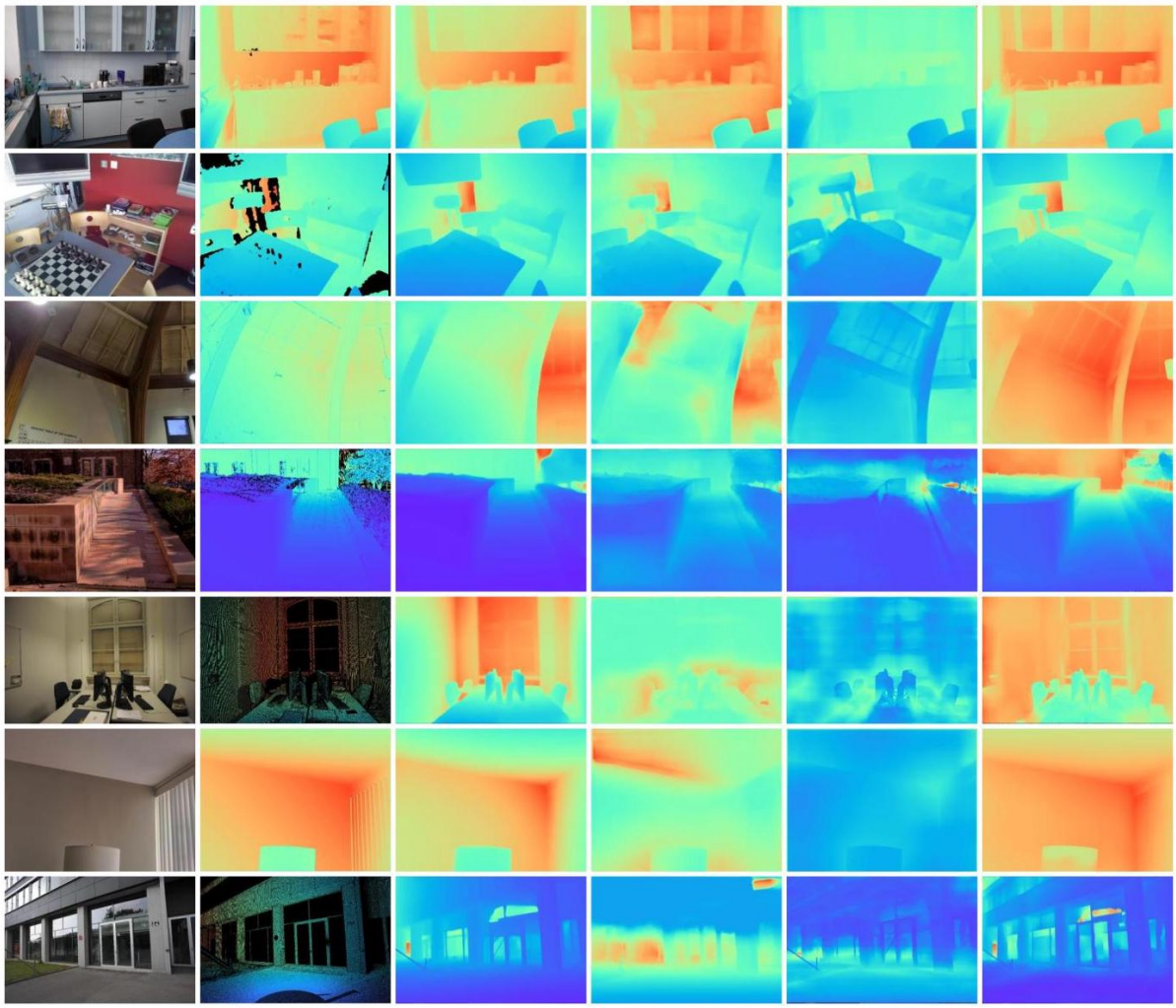

Figure 5: The visual comparison of predicted depth on iBims, ETH3D, and DIODE. Our depth maps come from the ConvNeXt-L CSTM_label model.

图 5: 对 iBims、ETH3D 和 DIODE 的预测深度进行可视化比较。我们的深度图来自 ConvNeXt - L CSTM_label 模型。

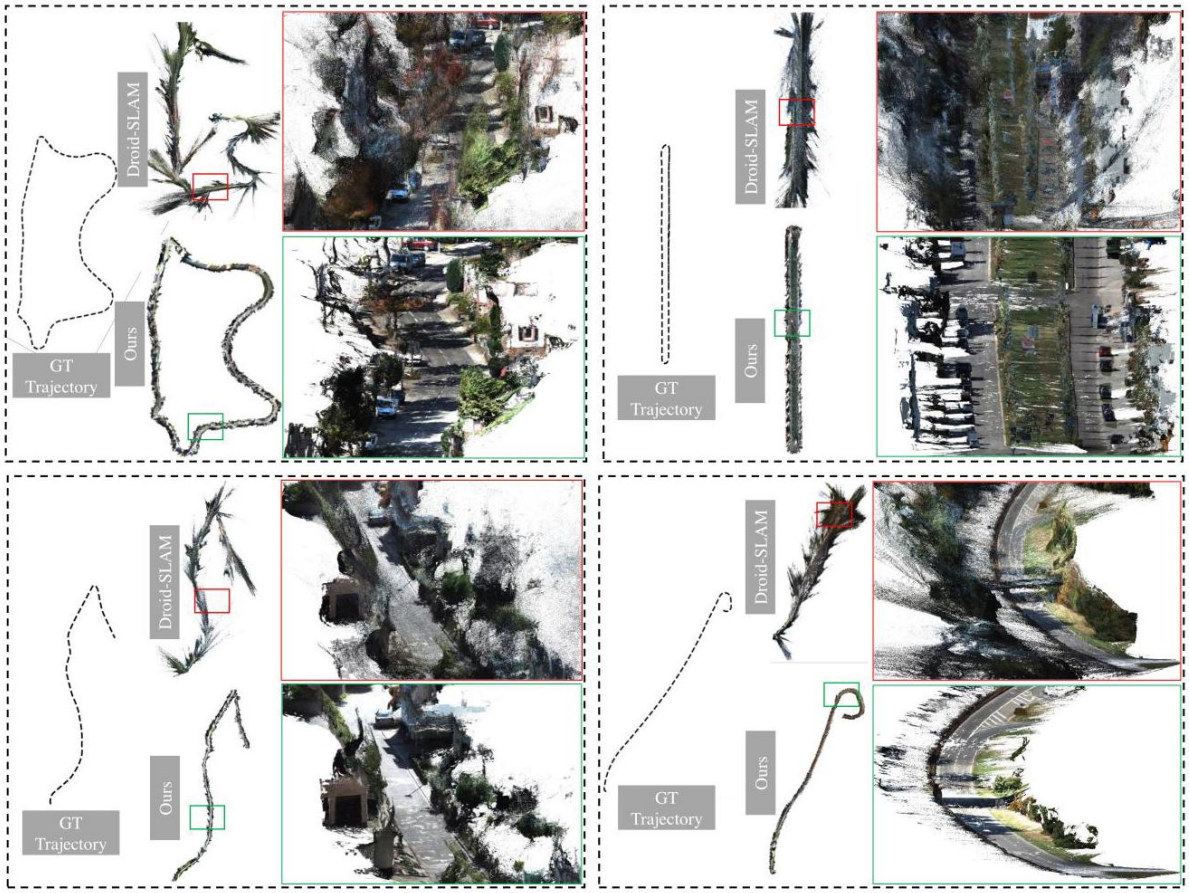

Figure 6: Dense-SLAM Mapping. Existing SOTA mono-SLAM methods usually face scale drift problems in large-scale scenes and are unable to achieve the metric scale. We show the ground-truth trajectory and Droid-SLAM [39] predicted trajectory and their dense mapping. Then, we naively input our metric depth to Droid-SLAM, which can recover a much more accurate trajectory and perform the metric dense mapping.

图 6: 密集同步定位与地图构建 (Dense - SLAM)。现有的最先进单目同步定位与地图构建 (mono - SLAM) 方法在大规模场景中通常会面临尺度漂移问题，无法实现度量尺度。我们展示了真实轨迹和 Droid - SLAM [39] 预测的轨迹及其密集地图。然后，我们简单地将我们的度量深度输入到 Droid - SLAM 中，它可以恢复出更准确的轨迹并进行度量密集地图构建。

Figure 7: 3D metric reconstruction of in-the-wild images. We collect several Flickr images and use our model to reconstruct the scene. The focal length information is collected from the photo's metadata. From the reconstructed point cloud, we can measure some structures' sizes. We can observe that sizes are in a reasonable range.

图 7: 野外图像的 3D 度量重建。我们收集了几张 Flickr 图像，并使用我们的模型对场景进行重建。焦距信息从照片的元数据中获取。从重建的点云，我们可以测量一些结构的尺寸。我们可以观察到尺寸在合理范围内。

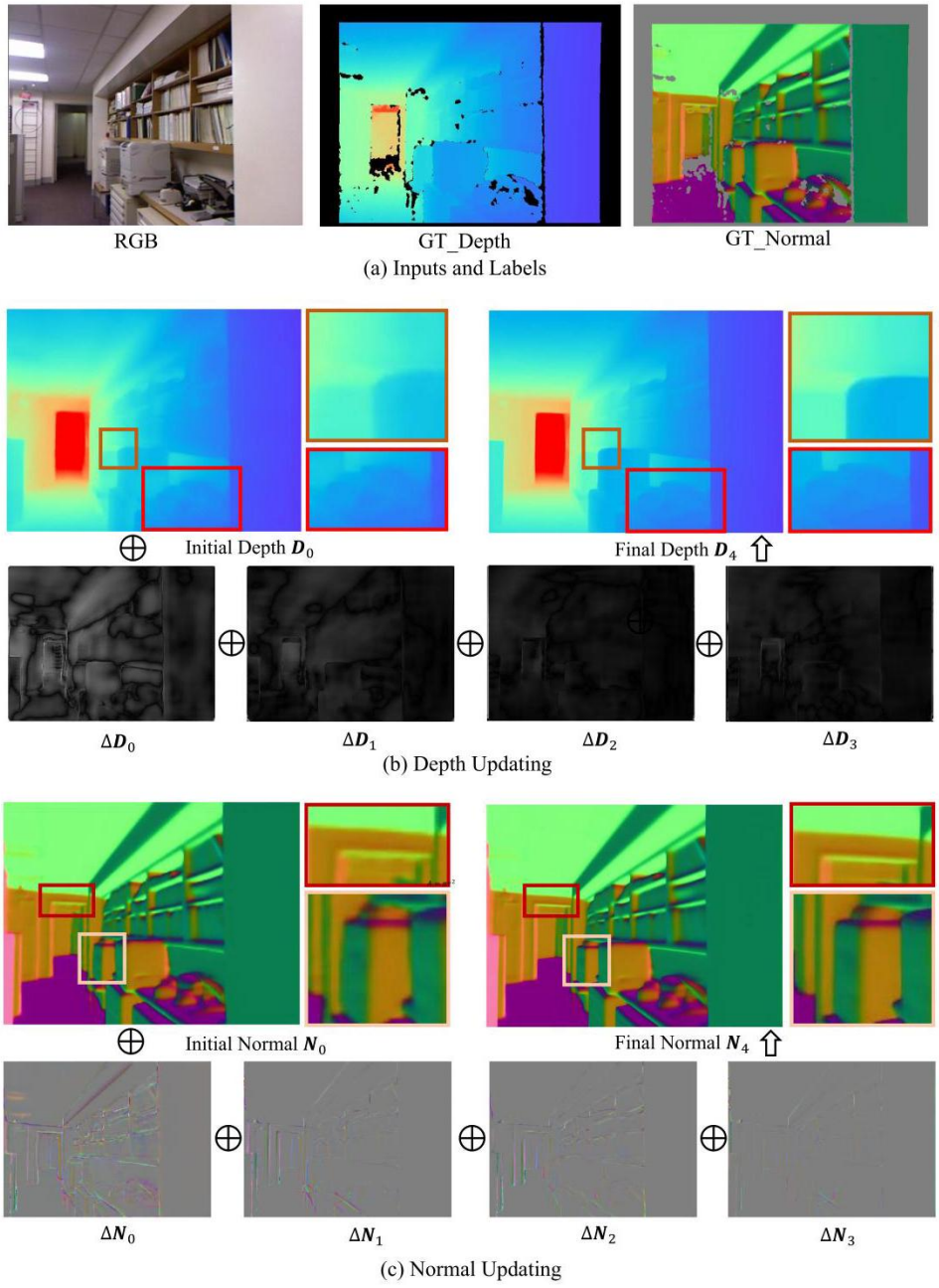

Figure 8: Visualization for iterative optimization. We use our publicly available ViT-S model (with 4 refinement steps) to estimate zero-shot depth and normal maps. The initial and final predictions, as well as their sequential updating items, are presented in (b) and (c).

图 8: 迭代优化可视化。我们使用我们公开可用的 ViT - S 模型 (有 4 个细化步骤) 来估计零样本深度图和法线图。初始和最终预测以及它们的顺序更新项分别在 (b) 和 (c) 中展示。

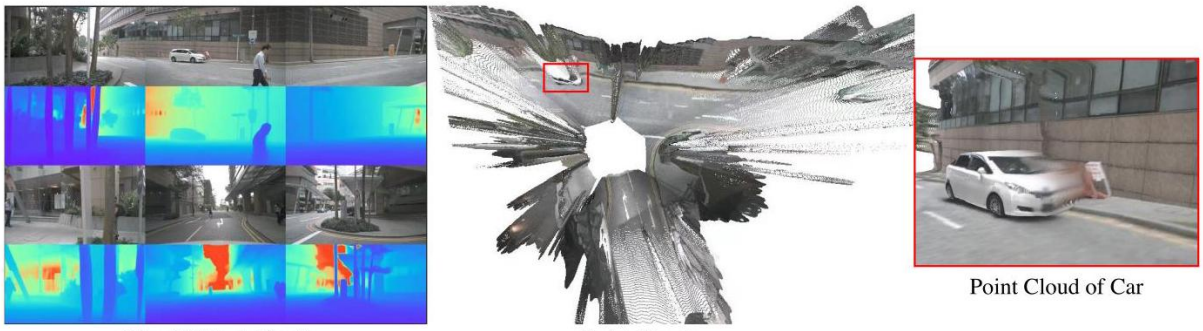

Figure 9: 3D reconstruction of 360° views. Current autonomous driving cars are equipped with several pinhole cameras to capture 360° views. With our model, we can reconstruct each view and smoothly fuse them together. We can see that all views can be well merged together without scale inconsistency problems. Testing data are from NuScenes. Note that the front view camera has a different focal length from other views.

图 9: 360° 视角的 3D 重建。当前的自动驾驶汽车配备了多个针孔相机来捕捉 360° 视角。使用我们的模型，我们可以重建每个视角并将它们平滑地融合在一起。我们可以看到所有视角都可以很好地融合在一起，没有尺度不一致的问题。测试数据来自 NuScenes。请注意，前视相机的焦距与其他视角不同。

[14] M. Gehrig, W. Aarents, D. Gehrig, and D. Scara-muzza, "Dsec: A stereo event camera dataset for driving scenarios," IEEE Robotics and Automation Letters, 2021.

M. 格里格 (M. Gehrig)、W. 阿伦茨 (W. Aarents)、D. 格里格 (D. Gehrig) 和 D. 斯卡拉穆扎 (D. Scara - muzza)，"Dsec: 用于驾驶场景的立体事件相机数据集"，《电气与电子工程师协会机器人与自动化快报》(IEEE Robotics and Automation Letters)，2021 年。

[15] M. Lopez-Antequera, P. Gargallo, M. Hofinger, S. R. Bulò, Y. Kuang, and P. Kontschieder, "Mapillary planet-scale depth dataset," in Proc. Eur. Conf. Comp. Vis., vol. 12347, pp. 589-604, 2020.

M. 洛佩斯 - 安特克拉 (M. Lopez - Antequera)、P. 加尔加洛 (P. Gargallo)、M. 霍芬格 (M. Hofinger)、S. R. 布洛 (S. R. Bulò)、Y. 匡 (Y. Kuang) 和 P. 孔奇德 (P. Kontschieder)，《Mapillary 行星尺度深度数据集》，载于《欧洲计算机视觉会议论文集》(Proc. Eur. Conf. Comp. Vis.)，第 12347 卷，第 589 - 604 页，2020 年。

[16] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, Y. Wang, and D. Yang, "Pandaset: Advanced sensor suite dataset for autonomous driving," in IEEE Int. Intelligent Transportation Systems Conf., 2021.

肖鹏 (P. Xiao)、邵志 (Z. Shao)、郝帅 (S. Hao)、张泽 (Z. Zhang)、柴鑫 (X. Chai)、焦健 (J. Jiao)、李泽 (Z. Li)、吴杰 (J. Wu)、孙凯 (K. Sun)、蒋凯 (K. Jiang)、王宇 (Y. Wang) 和杨迪 (D. Yang)，"Pandaset: 用于自动驾驶的先进传感器套件数据集"，收录于 2021 年 IEEE 国际智能交通系统会议论文集。

[17] Z. Bauer, F. Gomez-Donoso, E. Cruz, S. Orts-Escolano, and M. Cazorla, "Uasol, a large-scale high-resolution outdoor stereo dataset," Scientific data, vol. 6, no. 1, pp. 1-14, 2019.

Z. 鲍尔 (Z. Bauer)、F. 戈麦斯 - 多诺索 (F. Gomez-Donoso)、E. 克鲁兹 (E. Cruz)、S. 奥尔茨 - 埃斯科拉诺 (S. Orts-Escolano) 和 M. 卡索拉 (M. Cazorla),《Uasol: 一个大规模高分辨率户外立体数据集》,《科学数据》(Scientific data)，第 6 卷，第 1 期，第 1 - 14 页，2019 年。

[18] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," arXiv preprint arXiv:2001.10773, 2020.

Y. 卡邦 (Y. Cabon)、N. 默里 (N. Murray) 和 M. 胡门伯格 (M. Humenberger),"虚拟基蒂数据集 2(Virtual kitti 2)"，预印本 arXiv:2001.10773，2020 年。

[19] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., "Scalability in perception for autonomous driving: Waymo open dataset," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2446-2454, 2020.

孙 (Sun)、克雷奇马尔 (Kretzschmar)、多蒂瓦拉 (Dotiwalla)、舒阿尔 (Chouard)、帕特奈克 (Patnaik)、崔 (Tsui)、郭 (Guo)、周 (Zhou)、柴 (Chai)、凯恩 (Caine) 等人，"自动驾驶感知的可扩展性:Waymo 开放数据集"，收录于《电气与电子工程师协会/计算机视觉基金会计算机视觉与模式识别会议论文集》，第 2446 - 2454 页，2020 年。

[20] A. Zamir, A. Sax,, W. Shen, L. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task

A. 扎米尔 (A. Zamir)、A. 萨克斯 (A. Sax)、W. 沈 (W. Shen)、L. 吉巴斯 (L. Guibas)、J. 马利克 (J. Malik) 和 S. 萨瓦雷塞 (S. Savarese)，"任务分类学 (Taskonomy): 解开任务

transfer learning," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., IEEE, 2018.

迁移学习 (transfer learning)，见《IEEE 计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，IEEE，2018 年。

[21] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wi-jmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, et al., "The replica dataset: A digital replica of indoor spaces," arXiv preprint arXiv:1906.05797, 2019.

J. 斯特劳布 (J. Straub)、T. 惠兰 (T. Whelan)、L. 马 (L. Ma)、Y. 陈 (Y. Chen)、E. 维伊曼斯 (E. Wi-jmans)、S. 格林 (S. Green)、J. J. 恩格尔 (J. J. Engel)、R. 穆尔 - 阿塔尔 (R. Mur-Artal)、C. 任 (C. Ren)、S. 维尔马 (S. Verma) 等,《副本数据集: 室内空间的数字副本》，预印本 arXiv:1906.05797，2019 年。

[22] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3 d reconstructions of indoor scenes," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 5828-5839, 2017.

戴 (Dai)、张 (Chang)、萨瓦 (Savva)、哈尔伯 (Halber)、芬克豪泽 (Funkhouser) 和尼斯纳 (Nießner),"扫描网络 (ScanNet): 室内场景的丰富注释 3 d 重建"，载于《电气与电子工程师协会计算机视觉与模式识别会议论文集》，第 5828 - 5839 页，2017 年。

[23] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Under-sander, W. Galuba, A. Westbury, A. X. Chang, et al., "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," arXiv preprint arXiv:2109.08238, 2021.

S. K. 拉马克里什南 (S. K. Ramakrishnan)、A. 戈卡斯兰 (A. Gokaslan)、E. 维伊曼斯 (E. Wijmans)、O. 马克西梅茨 (O. Maksymets)、A. 克莱格 (A. Clegg)、J. 特纳 (J. Turner)、E. 安德桑德 (E. Under-sander)、W. 加卢巴 (W. Galuba)、A. 韦斯特伯里 (A. Westbury)、A. X. 张 (A. X. Chang) 等，“栖息地 - Matterport 3D 数据集 (Habitat-matterport 3D dataset，HM3D): 用于具身人工智能的 1000 个大规模 3D 环境”，预印本 arXiv:2109.08238，2021 年。

[24] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, ”Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding,” in Proceedings of the IEEE/CVF international conference on computer vision, pp. 10912-10922, 2021.

M. 罗伯茨 (M. Roberts)、J. 拉马普拉姆 (J. Ramapuram)、A. 兰詹 (A. Ranjan)、A. 库马尔 (A. Kumar)、M. A. 包蒂斯塔 (M. A. Bautista)、N. 帕赞 (N. Paczan)、R. 韦伯 (R. Webb) 和 J. M. 萨斯金德 (J. M. Susskind),《Hypersim: 用于整体室内场景理解的逼真合成数据集》, 载于《电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集》，第 10912 - 10922 页，2021 年。

[25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, ”Indoor segmentation and support inference from rgbd images,” in Proc. Eur. Conf. Comp. Vis., pp. 746-760, Springer, 2012.

N. 西尔伯曼 (N. Silberman)、D. 霍耶姆 (D. Hoiem)、P. 科利 (P. Kohli) 和 R. 弗格斯 (R. Fergus)，“基于 RGB - D 图像的室内分割与支撑推断”，收录于《欧洲计算机视觉会议论文集》(Proc. Eur. Conf. Comp. Vis.)，第 746 - 760 页，施普林格出版社 (Springer)，2012 年。

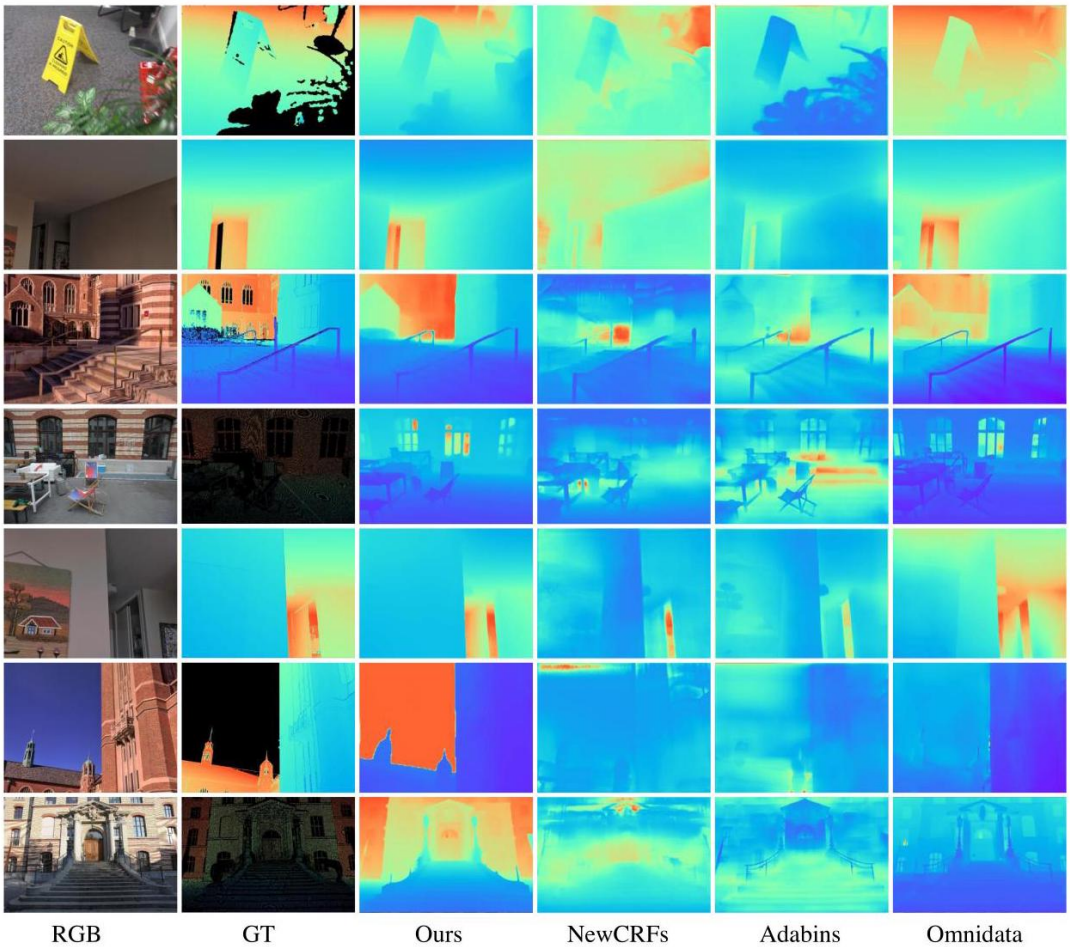

Figure 10: Depth estimation. The visual comparison of predicted depth on iBims, ETH3D, and DIODE. Our depth maps come from the ConvNeXt-L CSTM_label model.

图 10: 深度估计。iBims、ETH3D 和 DIODE 数据集上预测深度的可视化对比。我们的深度图来自 ConvNeXt - L CSTM_label 模型。

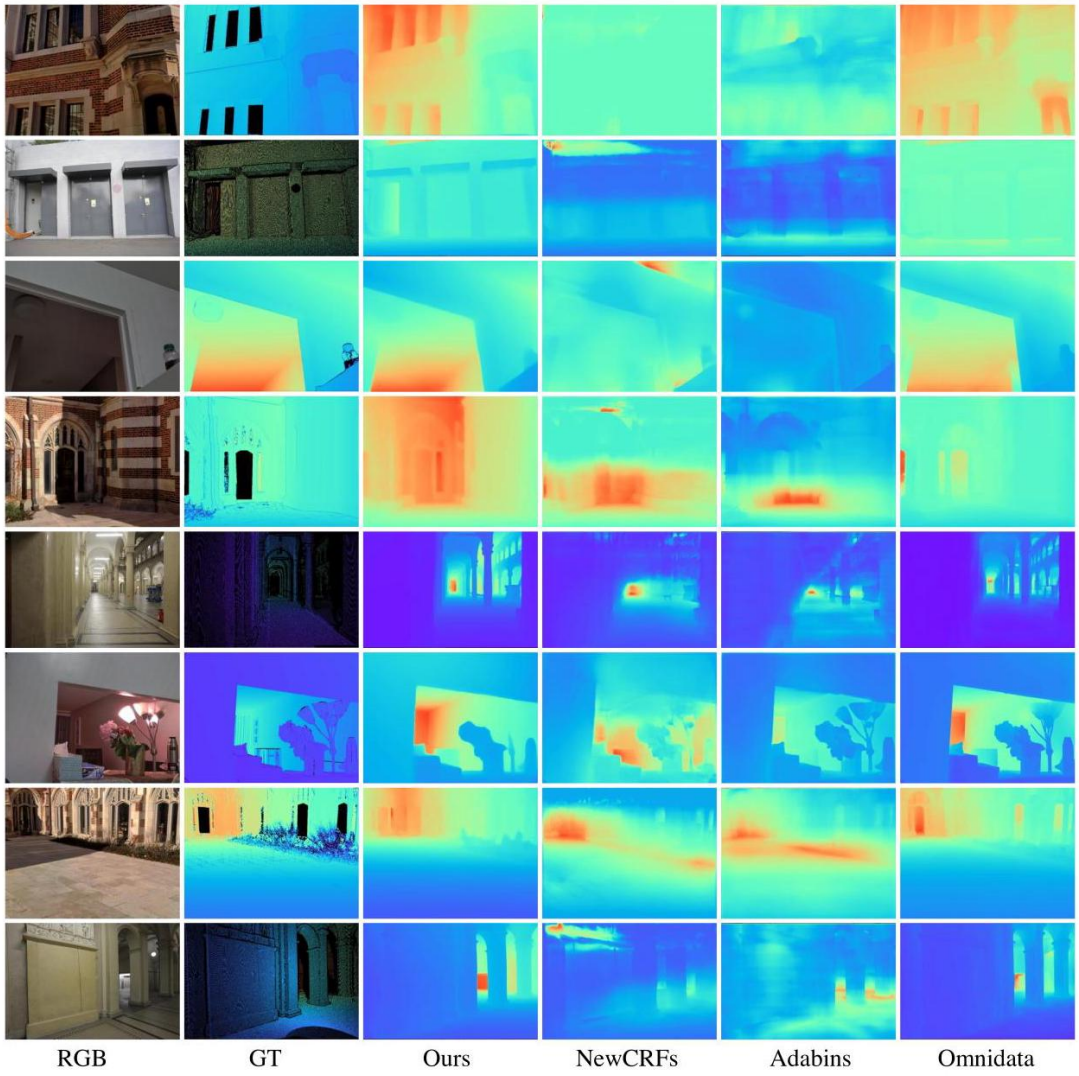

Figure 11: Depth estimation. The visual comparison of predicted depth on iBims, ETH3D, and DIODE. Our depth maps come from the ConvNeXt-L CSTM_label model.

图 11: 深度估计。iBims、ETH3D 和 DIODE 数据集上预测深度的可视化对比。我们的深度图来自 ConvNeXt - L CSTM_label 模型。

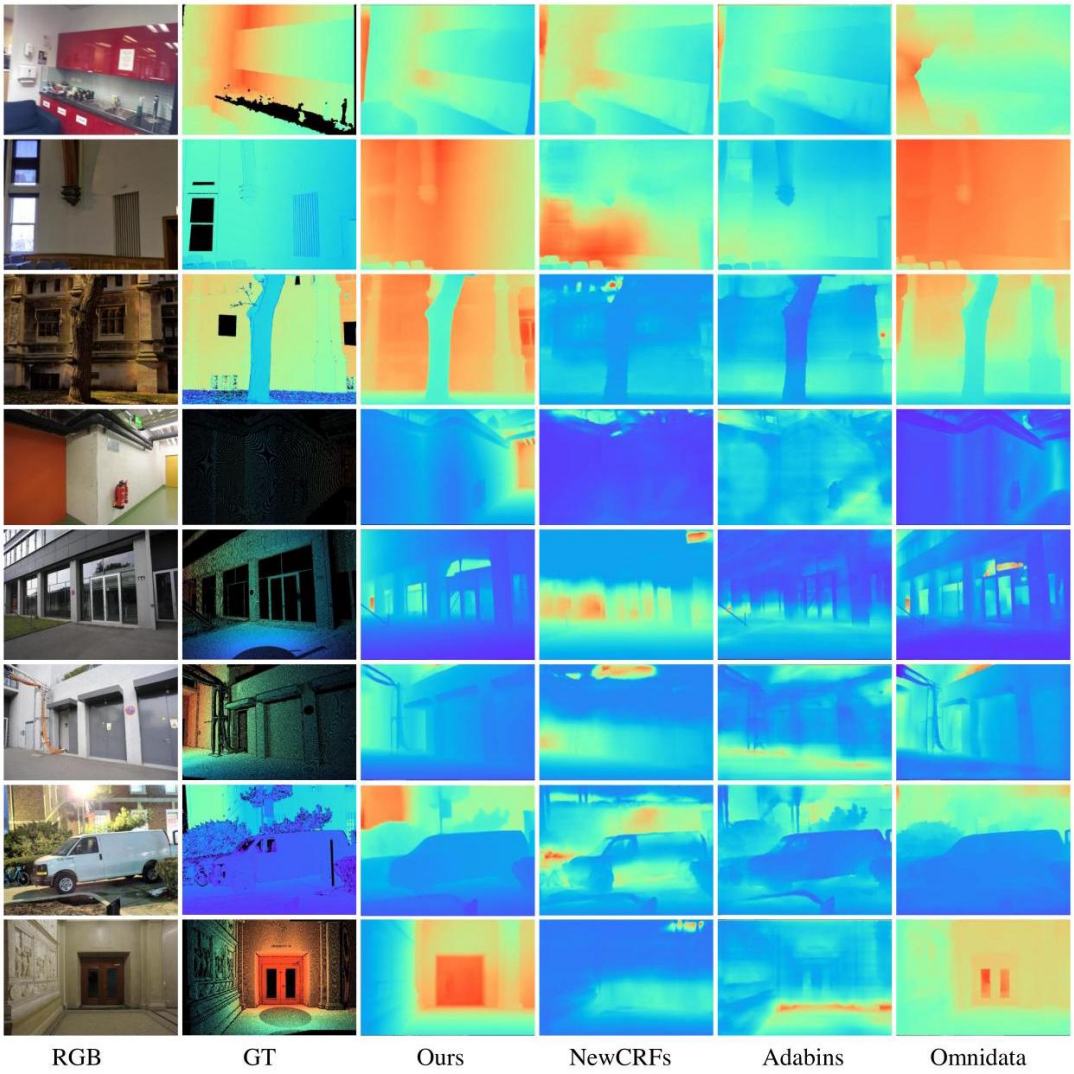

Figure 12: Depth estimation. The visual comparison of predicted depth on iBims, ETH3D, and DIODE. Our depth maps come from the ConvNeXt-L CSTM_label model.

图 12: 深度估计。iBims、ETH3D 和 DIODE 数据集上预测深度的可视化对比。我们的深度图来自 ConvNeXt - L CSTM_label 模型。

[26] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," Int. J. Robot. Res., 2013.

A. 盖格 (A. Geiger)、P. 伦茨 (P. Lenz)、C. 施蒂勒 (C. Stiller) 和 R. 乌尔塔松 (R. Urtasun)，“视觉与机器人技术的结合: 基蒂数据集 (The kitti dataset)”，《国际机器人研究杂志》(Int. J. Robot. Res.)，2013 年。

[27] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 11621-11631, 2020.

H. 凯撒 (H. Caesar)、V. 班基蒂 (V. Bankiti)、A. H. 朗 (A. H. Lang)、S. 沃拉 (S. Vora)、V. E. 利翁 (V. E. Liong)、Q. 徐 (Q. Xu)、A. 克里什南 (A. Krishnan)、Y. 潘 (Y. Pan)、G. 巴尔丹 (G. Baldan) 和 O. 贝伊博姆 (O. Beijbom)，“努斯场景数据集 (nuscenes): 用于自动驾驶的多模态数据集”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 11621 - 11631 页，2020 年。

[28] T. Schops, J. L. Schonberger, S. Galliani, T. Sat-tler, K. Schindler, M. Pollefeys, and A. Geiger, ”A multi-view stereo benchmark with high-resolution images and multi-camera videos,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 3260-3269, 2017.

T. 肖普斯 (T. Schops)、J. L. 舍恩贝格尔 (J. L. Schonberger)、S. 加利尼亚尼 (S. Galliani)、T. 萨特勒 (T. Sat - tler)、K. 辛德勒 (K. Schindler)、M. 波勒菲斯 (M. Pollefeys) 和 A. 盖格 (A. Geiger)，“具有高分辨率图像和多相机视频的多视图立体基准测试”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 3260 - 3269 页，2017 年。

[29] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, et al., ”Diode: A dense indoor and outdoor depth dataset,” arXiv: Comp. Res. Repository, p. 1908.00463, 2019.

I. 瓦西列维奇 (I. Vasiljevic)、N. 科尔金 (N. Kolkin)、S. 张 (S. Zhang)、R. 罗 (R. Luo)、H. 王 (H. Wang)、F. Z. 戴 (F. Z. Dai)、A. F. 丹尼尔 (A. F. Daniele)、M. 莫斯塔贾比 (M. Mostajabi)、S. 巴萨特 (S. Basart)、M. R. 沃尔特 (M. R. Walter) 等，“二极管数据集 (Diode): 一个密集的室内外深度数据集”，预印本服务器 (arXiv: Comp. Res. Repository)，第 1908.00463 页，2019 年。

[30] T. Koch, L. Liebel, F. Fraundorfer, and M. Korner, ”Evaluation of cnn-based single-image depth estimation methods,” in Eur. Conf. Comput. Vis. Worksh., pp. $0-0, 2018$.

T. 科赫 (T. Koch)、L. 利贝尔 (L. Liebel)、F. 弗劳恩多尔弗 (F. Fraundorfer) 和 M. 科纳 (M. Korner)，“基于卷积神经网络的单图像深度估计方法评估”，收录于《欧洲计算机视觉研讨会论文集》(Eur. Conf. Comput. Vis. Worksh.)，第 $0-0, 2018$ 页。

[31] W. Yin, J. Zhang, O. Wang, S. Niklaus, S. Chen, Y. Liu, and C. Shen, ”Towards accurate reconstruction of 3 d scene shape from a single monocular image,” IEEE Trans. Pattern Anal. Mach. Intell., 2022.

W. 尹 (W. Yin)、J. 张 (J. Zhang)、O. 王 (O. Wang)、S. 尼克劳斯 (S. Niklaus)、S. 陈 (S. Chen)、Y. 刘 (Y. Liu) 和 C. 沈 (C. Shen)，“从单目图像准确重建 3 d 场景形状”，《电气与电子工程师协会模式分析与机器智能汇刊》(IEEE Trans. Pattern Anal. Mach. Intell.)，2022 年。

[32] W. Yin, X. Wang, C. Shen, Y. Liu, Z. Tian, S. Xu, C. Sun, and D. Renyin, ”Diversedepth: Affine-invariant depth prediction using diverse data,” arXiv: Comp. Res. Repository, p. 2002.00569, 2020.

W. 尹 (W. Yin)、X. 王 (X. Wang)、C. 沈 (C. Shen)、Y. 刘 (Y. Liu)、Z. 田 (Z. Tian)、S. 徐 (S. Xu)、C. 孙 (C. Sun) 和 D. 任寅 (D. Renyin)，“多样深度 (Diversedepth): 利用多样数据进行仿射不变深度预测”，预印本服务器 (arXiv: Comp. Res. Repository)，第 2002.00569 页，2020 年。

[33] P. Besl and N. McKay, ”Method for registration of 3- d shapes,” in Sensor fusion IV: Control Paradigms and Data Structures, vol. 1611, pp. 586-606, Spie, 1992.

P. 贝斯尔 (P. Besl) 和 N. 麦凯 (N. McKay)，“三维形状配准方法”，收录于《传感器融合 IV: 控制范式与数据结构》(Sensor fusion IV: Control Paradigms and Data Structures)，第 1611 卷，第 586 - 606 页，国际光学工程学会 (Spie)，1992 年。

[34] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, and C. Shen, ”Learning to recover 3d scene shape from a single image,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2021.

W. 尹 (W. Yin)、J. 张 (J. Zhang)、O. 王 (O. Wang)、S. 尼克劳斯 (S. Niklaus)、L. 麦 (L. Mai)、S. 陈 (S. Chen) 和 C. 沈 (C. Shen)，“从单张图像学习恢复三维场景形状”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，2021 年。

[35] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, ”Zoedepth: Zero-shot transfer by combining relative and metric depth,” arXiv preprint arXiv:2302.12288, 2023.

S. F. 巴特 (S. F. Bhat)、R. 比尔克尔 (R. Birkl)、D. 沃夫克 (D. Wofk)、P. 翁卡 (P. Wonka) 和 M. 米勒 (M. Müller)，“佐伊深度 (Zoedepth): 通过结合相对深度和度量深度实现零样本迁移”，预印本 (arXiv preprint arXiv:2302.12288)，2023 年。

[36] G. Bae, I. Budvytis, and R. Cipolla, ”Estimating and exploiting the aleatoric uncertainty in surface normal estimation,” in Proceedings of the

G. 贝 (G. Bae)、I. 布德维蒂斯 (I. Budvytis) 和 R. 奇波拉 (R. Cipolla)，“估计并利用表面法线估计中的偶然不确定性”，收录于

IEEE/CVF International Conference on Computer Vision, pp. 13137-13146, 2021.

电气与电子工程师协会/计算机视觉基金会国际计算机视觉会议论文集 (Proceedings of the IEEE/CVF International Conference on Computer Vision)，第 13137 - 13146 页，2021 年。

[37] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, ”Om-nidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 10786- 10796, 2021.

A. 埃夫特卡尔 (A. Eftekhar)、A. 萨克斯 (A. Sax)、J. 马利克 (J. Malik) 和 A. 扎米尔 (A. Zamir)，“全数据 (Om - nidata): 一种从三维扫描制作多任务中级视觉数据集的可扩展管道”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，第 10786 - 10796 页，2021 年。

[38] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, ”New CRFs: Neural window fully-connected CRFs for monocular depth estimation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2022.

W. 袁 (W. Yuan)、X. 顾 (X. Gu)、Z. 戴 (Z. Dai)、S. 朱 (S. Zhu) 和 P. 谭 (P. Tan)，“新型条件随机场 (New CRFs): 用于单目深度估计的神经窗口全连接条件随机场”，收录于《电气与电子工程师协会计算机视觉与模式识别会议论文集》(Proc. IEEE Conf. Comp. Vis. Patt. Recogn.)，2022 年。

[39] Z. Teed and J. Deng, ”Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras,” vol. 34, pp. 16558-16569, 2021.

Z. 蒂德 (Z. Teed) 和 J. 邓 (J. Deng),“德罗伊德视觉同步定位与地图构建 (Droid - slam): 用于单目、立体和 RGB - D 相机的深度视觉同步定位与地图构建”，第 34 卷，第 16558 - 16569 页，2021 年。