1. **问题介绍**

近年来随着人工智能的发展，自动驾驶技术飞速发展。以特斯拉为首的视觉派抛弃激光雷达，只依赖于图像进行三维感知。作为纯视觉感知方案的基石任务，基于图像的三维目标检测天然存在长尾问题。模型很可能会对数据集中没见过的类别物体漏检，而这种漏检往往是致命的。重建出整个三维场景可以作为一种安全冗余，在三维目标检测失效的情况下依然可以实现避障。

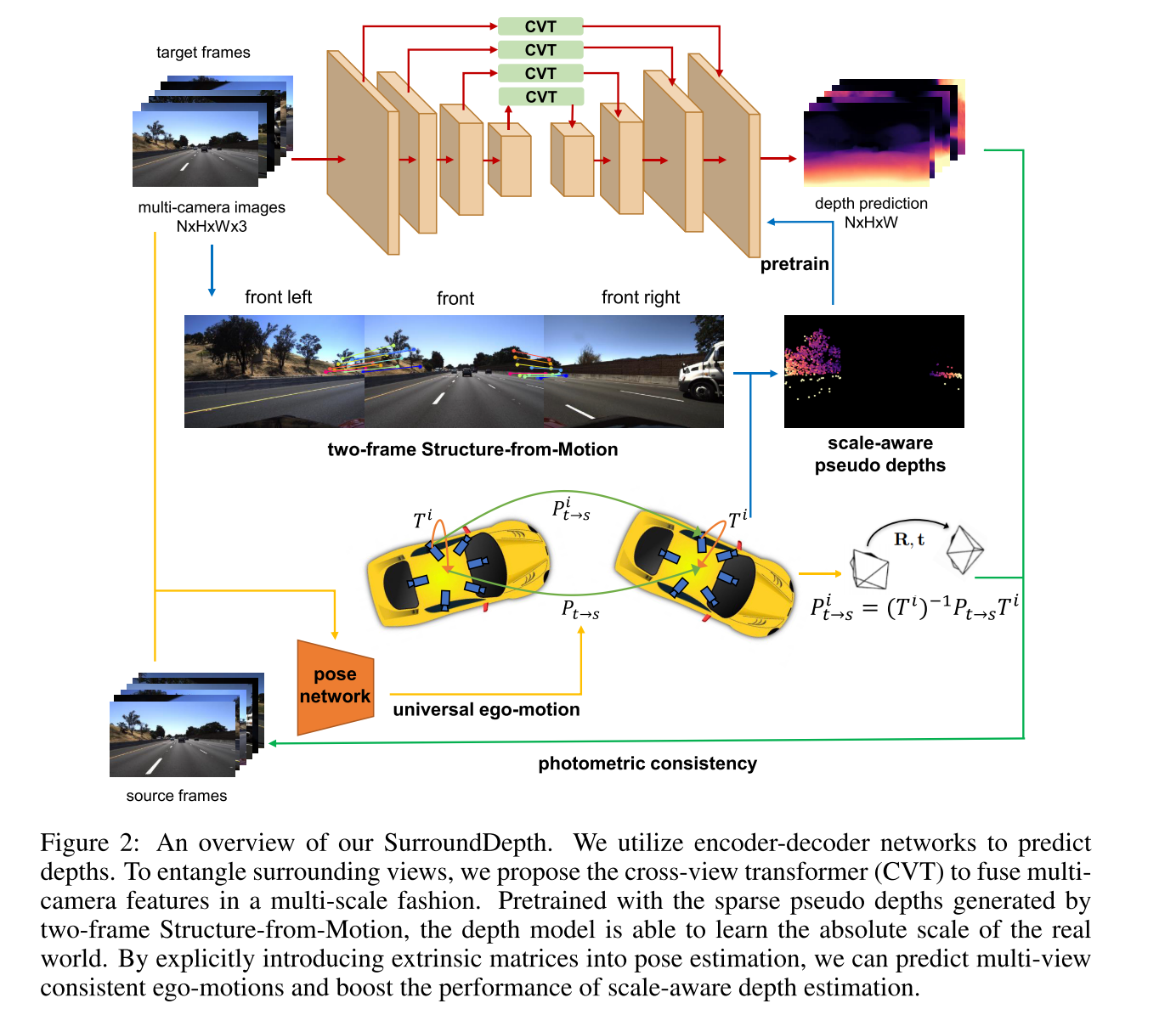
下面将介绍两篇此方面的论文，分别为自监督的SurroundDepth[1]和有监督的S3Depth[2]。

SurroundDepth：作为最简单直接且不需要点云标签的三维场景重建方式，在这个工作中我们重点研究自监督环视深度估计这个任务。 自监督深度估计是一个很经典的领域，早在17年就有相关的工作，但大部分工作都是基于单目图像的。与单目图像不同，环视图像的各个视角之间存在overlap，因此可以将多个视角之间的信息进行融合得到更准确的深度图预测。除此之外，自监督单目深度估计存在尺度歧义（scale-ambiguity）问题，换句话说，预测出的深度图会与深度真值差一个尺度系数。这是因为如果位姿和深度图同时乘以一个相同的尺度，会使得光度一致性误差（photometric loss）相同。文章作者根据环视视觉的特点提出了SurroundDepth，方法的核心是通过融合环视多视角信息以自监督的方式得到高精度且具有真实尺度的深度图。作者设计了跨视角transformer以注意力机制的形式对多视角的特征进行融合。为了恢复出真实尺度，笔者在相邻视角上利用SfM得到稀疏伪点云对网络进行预训练。之后提出联合位姿估计去显式地利用外参信息，对深度和位姿估计网络进行联合训练。在DDAD和nuScenes数据集上的实验验证了此方法超过了基线方法，达到了当前最佳性能。

S3Depth: 虽然在过去的几十年里，单目深度估计已经取得了很大的进展，但这些尝试主要是在只有前视相机的KITTI基准上进行的，忽略了周围视相机之间的相关性。在本文中提出了S3Depth，一种用于监督环视深度估计的简单基线，以联合预测多个周围相机的深度图。具体来说，采用了一个全局到局部的特征提取模块，该模块将CNN与transformer层相结合，以丰富表示。进一步，提出邻接视图注意机制，实现视图内和视图间的特征传播。前者由每个视图内的自注意模块实现，后者由相邻的注意模块实现，由相邻的注意模块计算跨多摄像机的注意力，交换环视图特征图的多尺度表示。大量的实验表明，我们的方法在DDAD和nuScenes数据集上都比现有的最先进的方法有更好的性能。

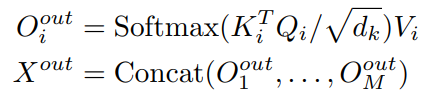
1. **相关求解算法简介及其实现**

2.1 SurroundDepth



**2.1.1跨视角Transformer （CVT）**

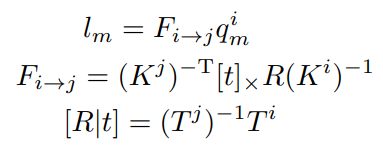
在获取所有视角的多尺度特征后，我们用多个CVT替代skip connection，连接encoder和decoder的同时，实现了多视角的特征交互，结构如下图所示。因为多个尺度的CVT操作是一致的，这里我们用一个尺度的CVT举例。假设第k个尺度的特征图可以表示为，在输入CVT之前，我们首先从视角（N），行（H）和列（W）三个维度，生成对应的positional encoding，即，为特征图提供位置信息。为了避免大尺度特征图导致的巨大计算量，我们利用常用的depthwise separable convolution降低大特征图的分辨率，同时保持特征通道数不变，可以表示为  。接下来我们介绍如何实现多视角的特征交互。首先，我们将多视角特征图展开成一个序列，序列的长度为N个特征图的分辨率之和，即  。然后，我们用3个线性层生成Q，K，V向量。最后、，我们采用经典的mult-head self-attention交换多视角特征，可以表示为：



除此之外，为了减少特征图下采样带来的信息丢失并保留输入细节，我们构造了一个跳跃连接来直接结合输入和输出特征：

**2.1.2 SfM预训练**

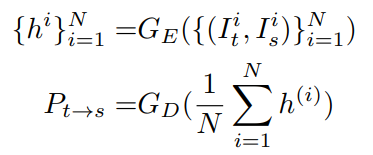
这一步的目的是为了挖掘外参包含的真实世界尺度信息。一个自然的做法是以外参作为位姿估计结果，在空域上利用photometric loss得到带有真实尺度的深度图。但环视图像之间的overlap比较小，这会使得在训练开始阶段，大部分的像素都会投影到overlap区域外，导致photometric loss无效，无法提供有效的真实尺度的监督。为了解决这个问题，我们用SIFT描述子对相邻视角的图像提取correspondences，并利用三角化转换成具有真实尺度的稀疏深度，并利用这些稀疏深度对深度估计网络进行预训练，使其可以预测带有真实尺度的深度图。但由于环视多视角之间的overlap较小，视角变化较大，因此描述子的准确度和鲁棒性有所降低。为了解决这个问题，我们首先预估出overlap区域，具体为每个视角图像左右1/3部分的图像，我们只在这些区域提取correspondences。进一步地，我们利用对极约束筛掉噪点：



其中为correspondence点对，为极线，为基本矩阵， 为第i个相机的内外参。若到距离大于某个阈值，我们认为这对correspondence是噪声，会将其滤掉。

**2.1.3 联合位姿估计**

大部分深度估计方法用PoseNet估计时序上相邻两帧的位姿。拓展到环视深度估计上，一个直接的方法是对每个视角单独预测位姿。但这种方法没有利用视角之间的几何变化关系，因此无法保证位姿之间的多视角一致性。为了解决这个问题，我们将位姿估计分解为两块。首先我们预测全局位姿，具体而言，我们将所有视角图像送入PoseNet encoder，将特征平均之后再送入decoder：



得到全局位姿 后，我们可以根据各个视角的相机外参进一步将其转化为每个相机的位姿：****其中  是第i 个相机的外参。由于联合位姿估计利用了含有真实尺度的外参，这一步有助于预测具有真实尺度的深度图。

2.2 S3Depth

**2.2.1 Problem Formulation**

给定N个视图输入样本I = {I1, I2，···IN}，期望包围视图深度模型f(·)预测深度图Dpred = {Dpred 1, Dpred 2，···Dpred N}，其中Ii∈R^H×W×3表示包围视图RGB图像，Dpred I∈R^H×W表示预测深度图，H, W分别表示输入图像的高度和宽度。在S3Depth中，唯一的监督信号来自激光雷达。与单目深度估计不同，环视图像之间存在重要的重叠，这是非常有用的，可以导入模型以整体理解场景。在此基础上，我们引入了一个简单的基线S3Depth来预测360◦视图深度的同时。S3Depth利用邻接视图信息的交互，可以获得比视相关估计更好的每个视图的性能。

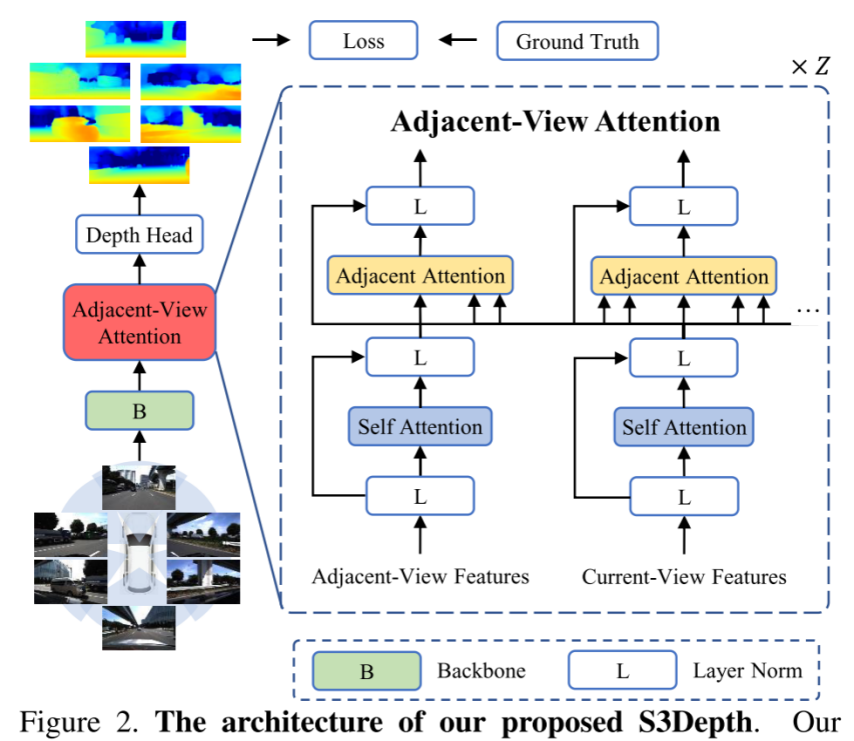
**2.2.2 Overview**

下图显示了我们建议的S3Depth架构。目的是建立一个强大的基线来预测环绕视图图像的密集深度，因此我们采用了一个多级模型来获得丰富的表示。其中，基线由三部分组成:特征提取、邻接视图注意和深度头。

(1)特征提取是一个共享的骨干，从输入图像的N个视图中提取2D特征

(2)利用邻接视图注意架构交换特征的邻接视图信息。具体来说，采用交替注意机制:自注意计算对同一视图中特征的注意，邻接注意计算对相邻视图中特征的注意，从而在每个尺度上交换相邻多尺度表示。

(3)与单目深度估计不同的是，深度头同时预测了环视深度图。



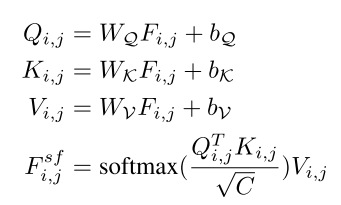
**2.2.3 Feature Extractor**

**Image Encoder:** 最近，注意力机制已经被证明在各种计算机视觉任务中是有效的，如图像分类，目标检测和深度估计。受益于电感偏置，transformer允许模型聚焦于特性的重要部分。尽管转换器可以为全局上下文感知建模，但它可能会忽略局部关系。相反，CNN可以利用当地的连通性。因此，我们在S3Depth中采用了一个结合CNN和transformer的模块，如图3所示，它可以学习局部和全局特征之间的交互，在五个阶段进行丰富的表示。其中，第一阶段采用convc -stem块提取2D特征F1∈R^N×C,F×H 2 ×W 2，其中H和W表示高度和图片的宽度。

**Depth Decoder:** 在深度解码器中，我们融合同一视图的多尺度特征和交叉视图图像的特征，逐步提高空间分辨率。然后分别在1处聚合特征。将1/2、1/4、1 /8、1/16个分辨率的输出映射到1/2、1/4、1/8、1/16个分辨率上采样到输入分辨率，然后计算损失函数。最后，利用跳跃连接连接输入和输出特征，减少下采样特征映射的信息损失。

**2.2.4 Adjacent-view Attention**

图2也提供了邻接视图注意体系结构的详细结构。与SurroundDepth不同，我们采用的是自我注意和邻接注意交替进行的方式。自注意提取当前视图中的特征，邻接注意用于在相邻视图特征之间交换表示。设Fij∈RN×CF H n ×W n, i = 1、2、3、4、5,j = 1、2、3、4、5、6为第i级第j个视图得到的特征图，其中H、W为图像的高度和宽度，n为输入图像的视图，CF为特征通道数。对于每个阶段，我们都使用了多头注意力来增加特征的表现力。注意机制利用点积相似度计算查询向量与关键向量之间的注意，然后对值向量进行加权。对于self-attention，使用特征描述符Fij作为输入，从同一个视图计算Q、K、V，其中WQ、WK、WV∈RC×C, bQ、bK、bV∈RC和C分别表示嵌入维数:



对于邻接注意，K、V由current-view特征计算，Q由邻接视图特征计算，其中x表示相邻视图(x∈j−1,j + 1):

**2.2.5 Depth Head**

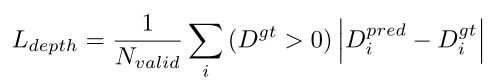
为了获得全面的理解和强大的表达能力，将同一位置不同尺度上的特征结合起来预测最终深度。通过利用邻接视图的优势，特征可以与邻接视图交互，以补偿错误的估计。设Fout表示Depth解码器和邻接注意的特征，最终输出可表示为:



σ为sigmoid激活函数，Dmax为最大距离。

**2.2.6 Loss Function**

我们直接使用点云数据作为深度估计的监督。具体来说，将点云投影到图像平面上。采用L1损耗计算有效像素与地面真值深度的距离，如下所示:

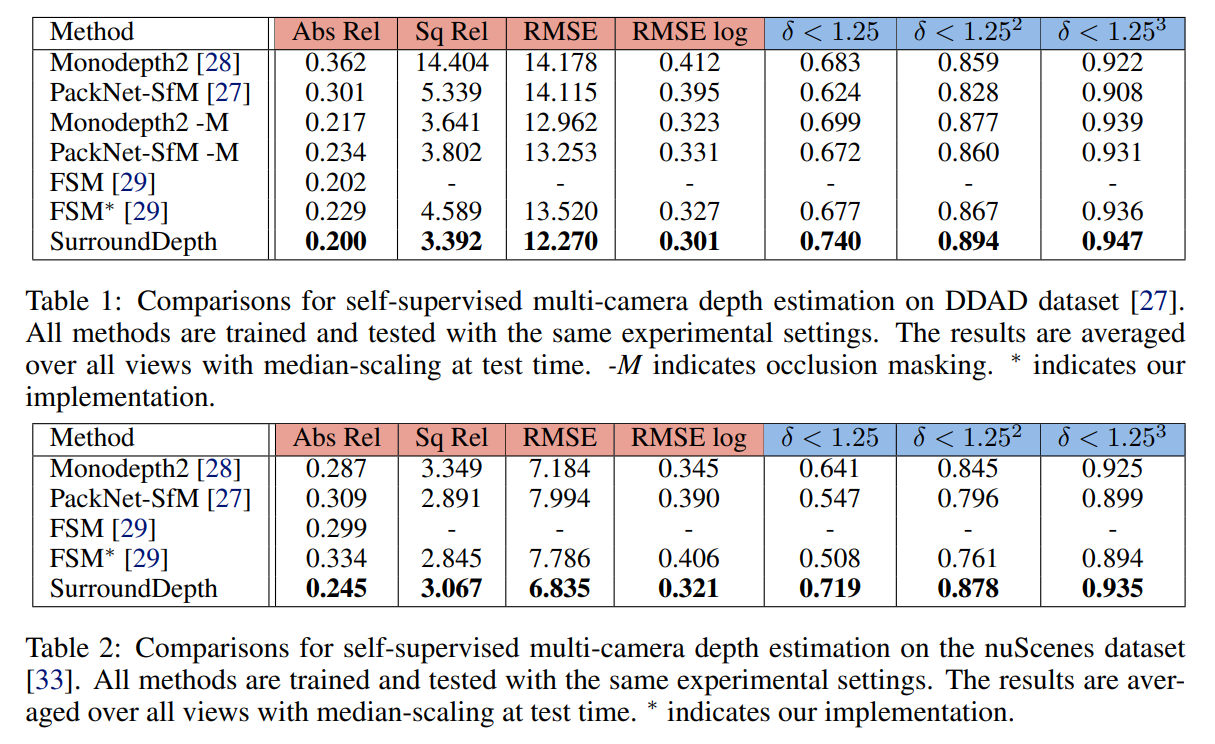


1. **最新发展、数据集、SOTA结果、实际运行结果等**

**3.1 数据集、最新发展（SOTA结果）**

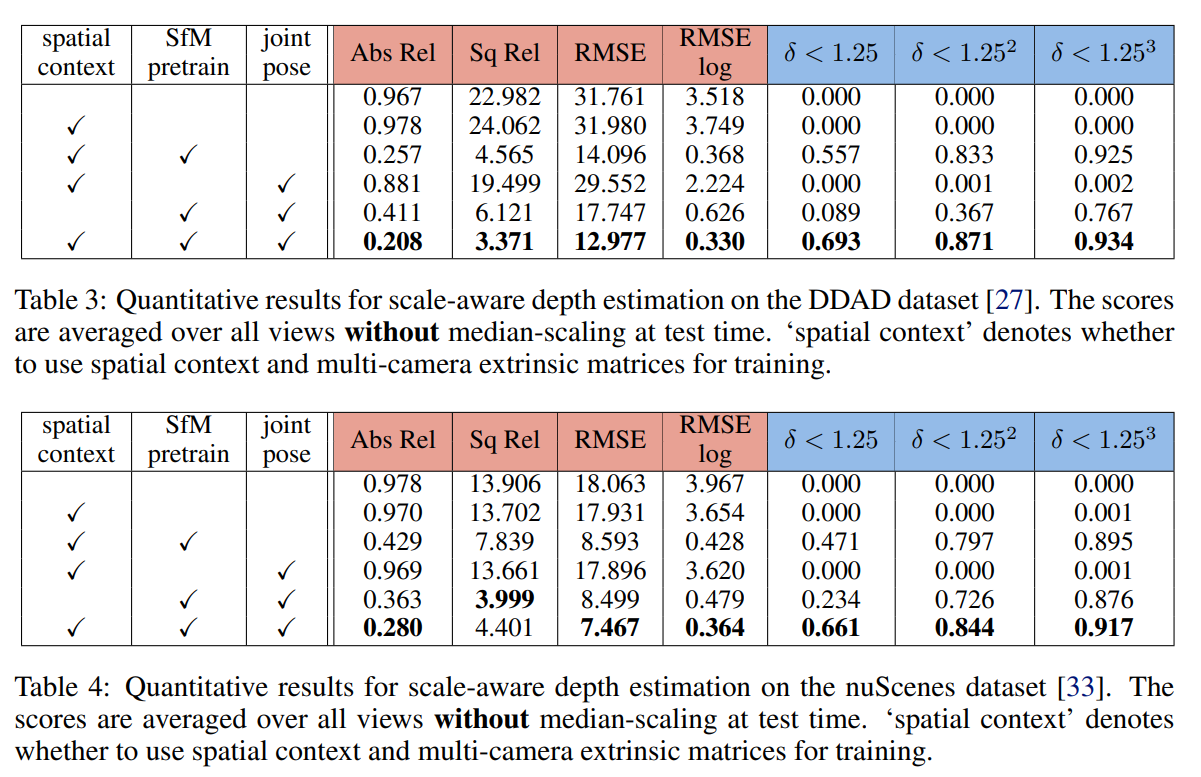
**surroundDepth：**

我们在DDAD(Dense Depth for Automated Driving) 以及nuScenes 上均进行了实验，使用了与Monodepth2 相同的backbone网络（ImageNet pretrained ResNet34）与pose estimation网络来构建SurroundDepth。在两个数据集上的实验结果如下：

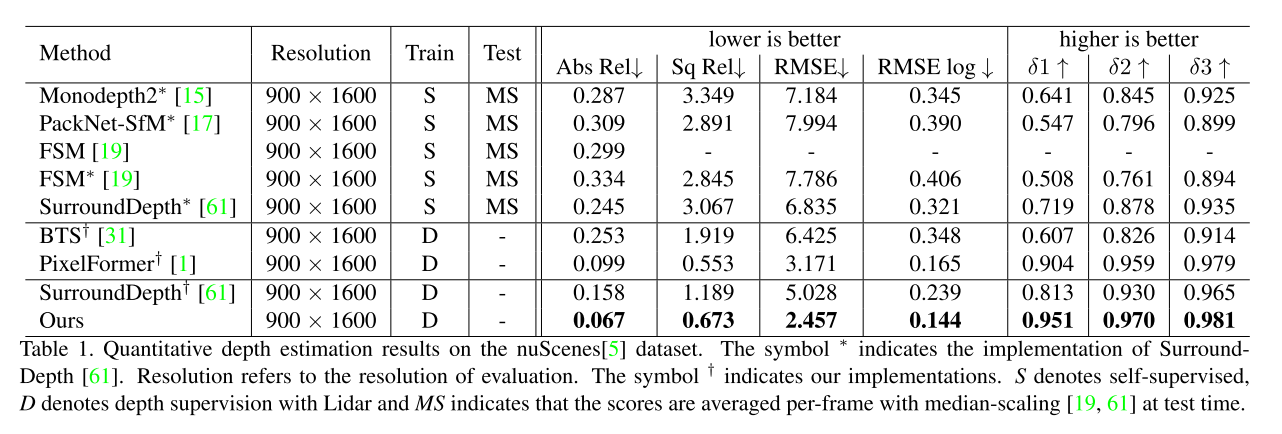


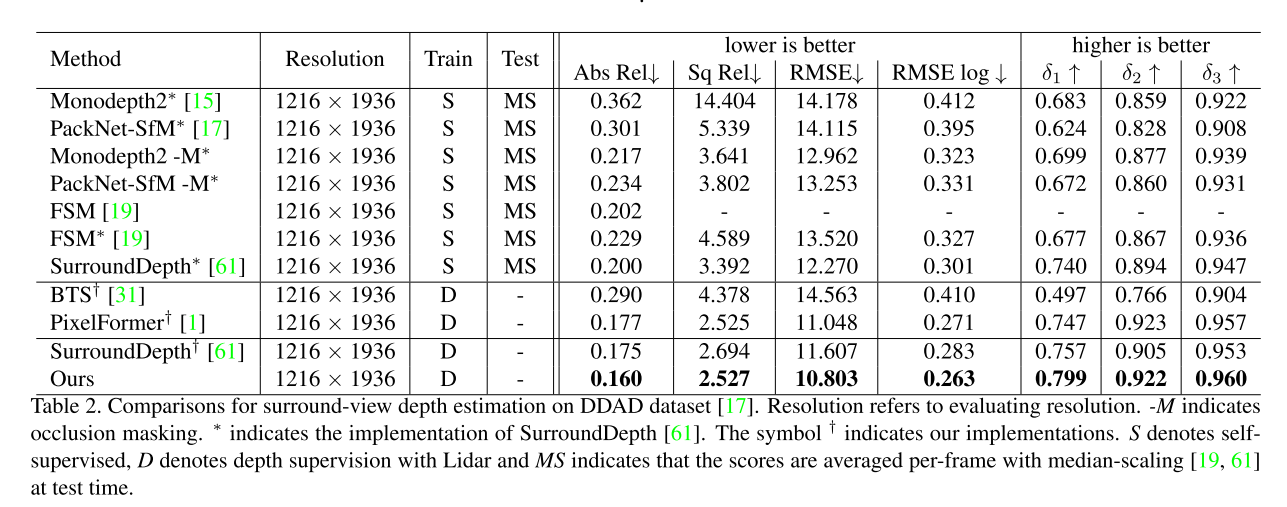
其中，我们对比了两种最先进的单目深度估计方法（Monodepth2 and PackNet-SfM ）以及一种多相机深度估计方法FSM 。我们在相同的测试环境下对比了所有的方法，可以看出，SurroundDepth在两个数据集上均取得了最好的性能。

此外，由于利用了环视相机之间的交互，SurroundDepth相比其他方法的一大优势在于可以取得绝对深度估计。针对绝对深度，我们在两个数据集上进行了相关实验。可以发现，仅仅利用spatial photometric loss无法使网络学习到绝对深度。通过我们提出的SfM pretraining方法，网络才能有效地预测绝对深度。

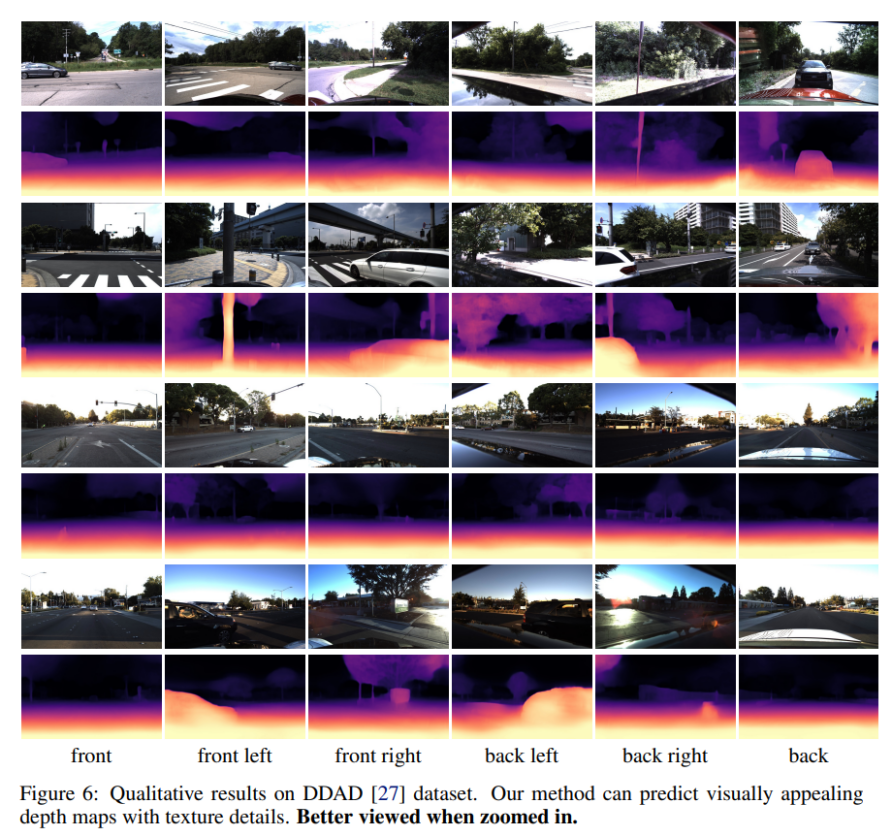
****

S3D的实验结果：

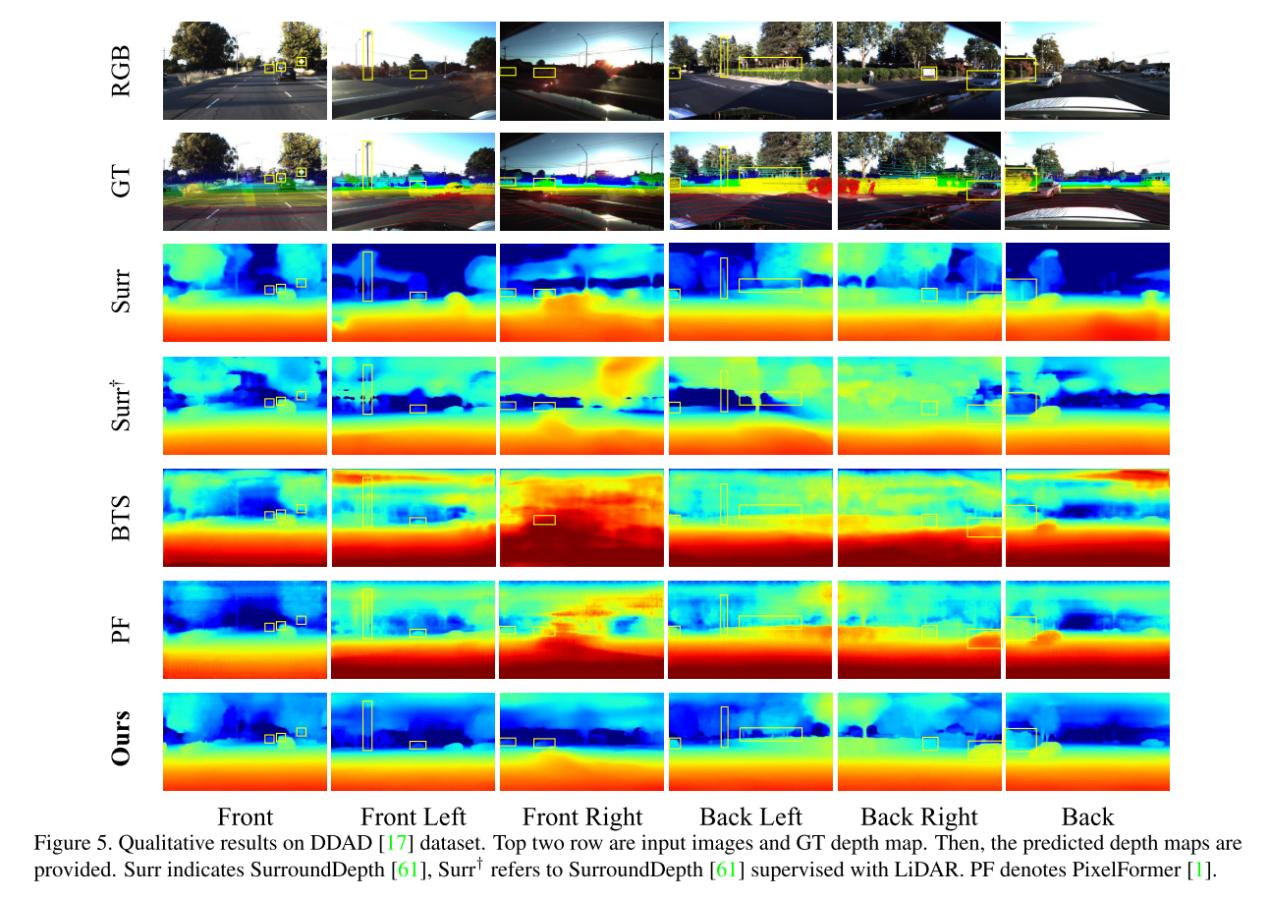




SurroundDepth在DDAD上的可视化结果如下图：



S3Ddpth的可视化结果：



1. **结论(conclusions)**

提出了 SurroundDepth，这是一种用于自监督多相机深度估计的方法。我们方法的核心观点是将多相机信息交织在一起，并共同处理所有周围的视图。交叉视图变换器在多个尺度上执行，以整合多视图特征。为了获得尺度感知的深度预测，我们提出了基于结构光从运动的预训练和联合姿态估计，以充分利用多相机外部矩阵。我们的方法在多相机深度估计数据集上实现了最先进的性能。

建立了一个简单的基线监督环视深度估计(S3Depth)。与单目深度估计不同，S3Depth的核心洞察力是交换相邻视图信息，共同预测周围视图的所有深度。我们使用了一个强大的主干，它将CNN与变压器层结合起来，以丰富表示。此外，在多个不同的尺度上，利用邻接视图的关注来整合邻接视图的特征。通过大量的实验，我们表明S3Depth可以在多摄像机深度估计数据集上实现最先进的性能。我们希望所提出的方法可以作为一个简单而强的基线多相机深度估计。

1. **参考文献(references)**

[1]Wei, Yi, et al. "SurroundDepth: entangling surrounding views for self-supervised multi-camera depth estimation." Conference on Robot Learning. PMLR, 2023.

[2]Guo X, Yuan W, Zhang Y, et al. A Simple Baseline for Supervised Surround-view Depth Estimation[J]. arXiv preprint arXiv:2303.07759, 2023.