# 《先进模式识别》课程报告

姓名：刘建

学号：22B903037

学院：计算学部

2023年6月

目录

[《先进模式识别》课程报告 1](#_Toc2125843932)

[1. 问题背景 3](#_Toc1373529985)

[2. 文献综述 4](#_Toc1616736292)

[2.1 MonoScene: Monocular 3D Semantic Scene Completion，CVPR 2022 4](#_Toc376274153)

[2.2 OccDepth: A Depth-aware Method for 3D Semantic Occupancy Network 7](#_Toc1847832703)

[2.3 VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion，CVPR 2023 10](#_Toc1763220054)

[2.4 TPVFormer: Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction，CVPR 2023 12](#_Toc1312602625)

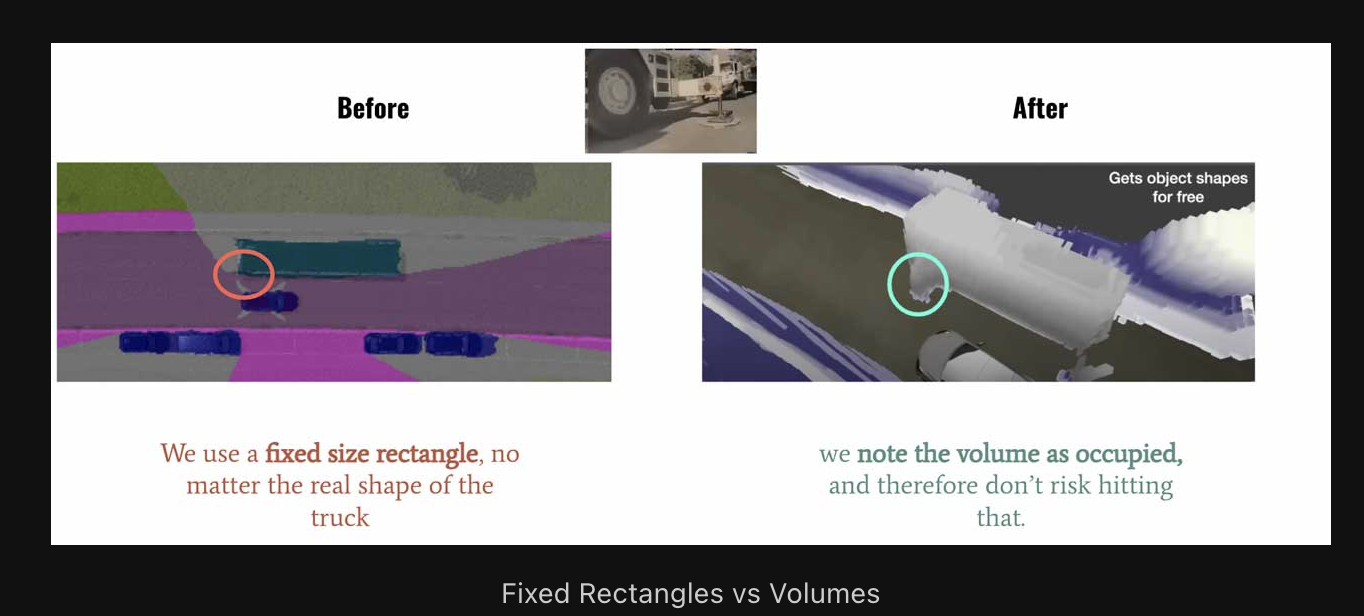
[3. 研究工作 15](#_Toc1960296391)

[4. 参考文献 17](#_Toc7331263)

# 问题背景

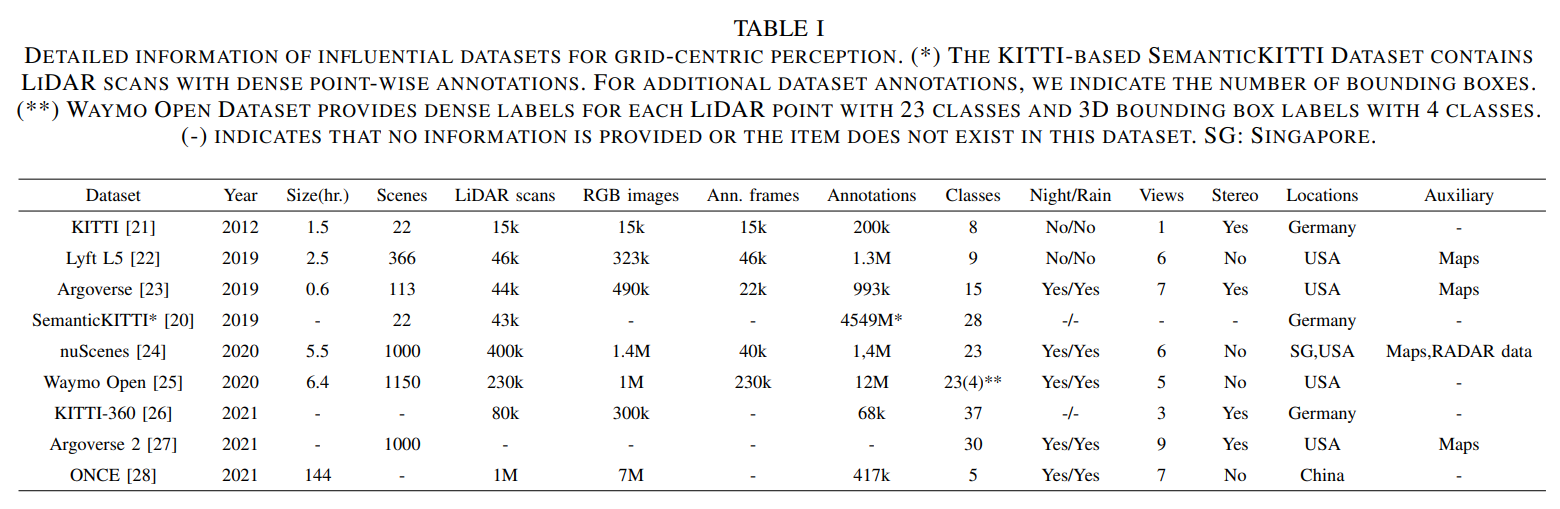
在 2022年的 Tesla AI Day上， Tesla 将Bev（鸟瞰图）感知进⼀步升级，提出了基于 Occupancy Network 的感知⽅法。这种基于Occupancy Grid Mappig的表示⽅法，又叫体素(Voxel)占据，在3D重建任务中已经是一个“老熟人”了。

它将世界划分成为⼀系列3D⽹格单元，然后定义哪个单元被占⽤，哪个单元是空闲的，并且每个占据单元同时也包含分类信息，⽐如路⾯、车辆、建筑物、树⽊等。在⾃动驾驶感知中，相⽐普通的3D检测⽅法，这种基于体素的表示可以帮助预测更精细的异形物体。如下图 Tesla Demo 中所展示的那样，对于空间感知更精细。

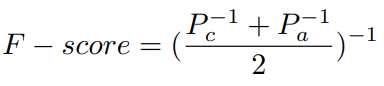


在这种在线重建的⽅法中，⼀般使⽤SSC ( Semantic Scene Completion)任务评判预测的准确性，即利⽤图像、点云或者其他3D数据作为输⼊，预测空间中的体素占据和类别信息，并与 GT标注相⽐较。

使用**数据集**如下：



**预测指标**：语义场景补全的主要度量是所有语义类的mIoU，在场景补全时使用IoU、Precision和Recall来评估几何重建质量。3D占用预测挑战测量F得分作为完整性Pc的调和平均值，精度Pa，F得分计算如下：



# 文献综述

本文调研了4篇基于视觉的3D Occupancy预测及语义补全的方法。该任务输入一帧或多帧图像，输出栅格形式的占据情况及语义类别。更多方法可以参考Grid-Centric的感知方法[1]。

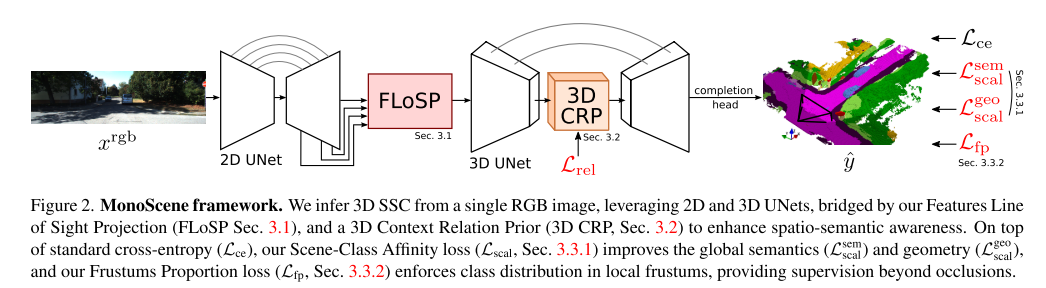
## 2.1 MonoScene: Monocular 3D Semantic Scene Completion，CVPR 2022

MonoScen[2]的主要贡献：提出了一种从单目图像直接预测三维语义栅格+语义补全的框架

#### 2.1.1 算法整体流程

方法的大致流程比较简单：

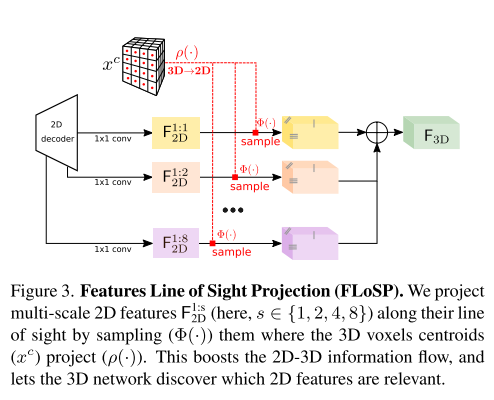
1. 2D UNet提取图像特征
2. 三维栅格从图像中提取特征
3. 类似3D Unet的结构提取三维特征
4. 预测head输出每个栅格是否占据+语义类别



#### 2.1.2 图像特征提取及投影

**图像特征提取：**采用预训练的EffecientNetB7，参数不固定，网络训练时也参与训练。对应代码monoscene/models/unet2d.py

**特征投影：**栅格的中心投影到图像平面上，在特征图1,2,4,8四个层级上插值得到对应的特征（对应代码monoscene/models/flosp.py），四个层级的特征求和后作为栅格特征（对应代码monoscene/models/monoscene.py）。超出图像范围的栅格特征设为0。

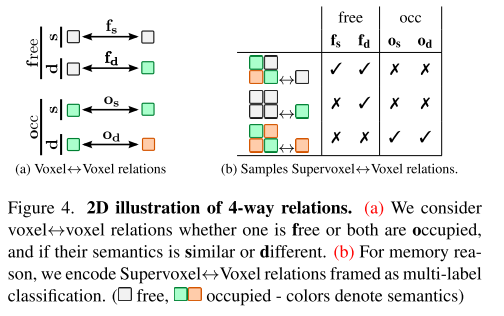


#### 2.1.3 三维特征提取及输出

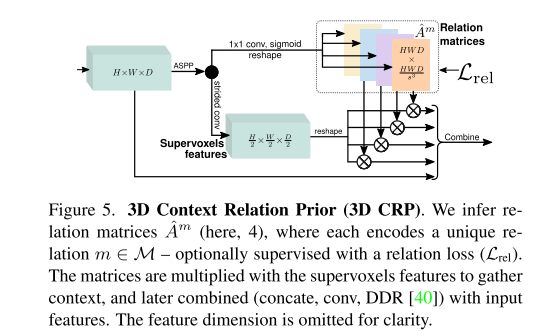
采用两级的3D UNet结构，其中最底层增加了一个3D Context Relation Prior（CRP）模块。之后接一个3D ASPP层+softmax输出占据和类别预测。这部分对应代码monoscene/models/unet3d\_xxx.py

**3D CRP** 模块主要的作用是预测每一个栅格和其他所有栅格是否属于同一类别，并分别把同类/不同类的其他栅格特征通过某种方式（类似加权求和）拿过来，和当前特征concat，作为最终特征。可以理解为带有inductive bias的cross attention。

**具体细节：**借鉴了CPNet这篇论文，建立一个(Z\*Y\*X)× (Z\*Y\*X/s^3)的关系矩阵描述voxel<->super voxel的关系（为了降低复杂度，将边长为s内的所有voxel合在一起考虑）。关系分为4种情况：（当前voxel和该super voxel均为空&语义类别不同），（当前voxel和该super voxel均为空&语义类别相同），（当前voxel和该super voxel有一个不空&语义类别相同），（当前voxel和该super voxel有一个不空&语义类别不同）。实际上只有三种情况，因为不存在voxel都为空且语义不相同的情况。



网络需要预测这个关系矩阵，通过关系矩阵可以得到4个attention map。关系矩阵和3D栅格特征（reshape到(Z/s\*Y/s\*X/s) × feature\_dim）矩阵相乘得到4个特征，通过concat等方式和当前特征组合，作为最终的特征。这部分对应代码monoscene/models/CRP3D.py（下图为s=2的示意图）

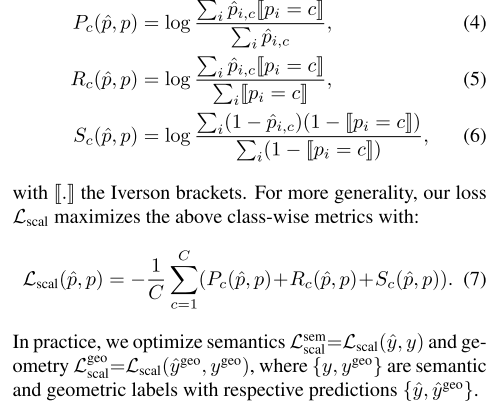


#### 2.1.4 训练

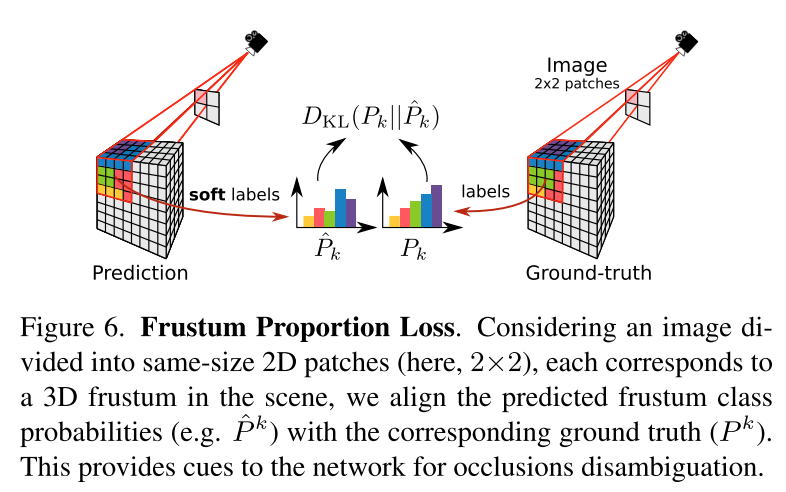
训练的loss包含4个部分：

* 每个栅格预测的占据/语义的Cross-entropy Loss，占据和语义合在一起，即free作为一个语义类别与语义的unknown合并为一类，其中free的类别权重为0.05
* 3D CRP预测的关系矩阵Cross-entropyLoss
* 3D CRP预测的Scene-Class Affinity Loss
* 作者提出的Frustum Proportion Loss

**Scene-Class Affinity Loss** 将类内voxel和类间voxel视为两个整体来指导网络学习关系矩阵，按照类似Percision, Recall, Specificity给出了loss的计算。



**Frustum Proportion Loss** 主要是为了让网络关注到视锥上类别的分布，比如视锥中一般会有地面（即使被物体遮挡），loss可以理解为一组视锥上语义真值的概率分布（栅格数量的比例）与网络实际预测的分布的KL散度。



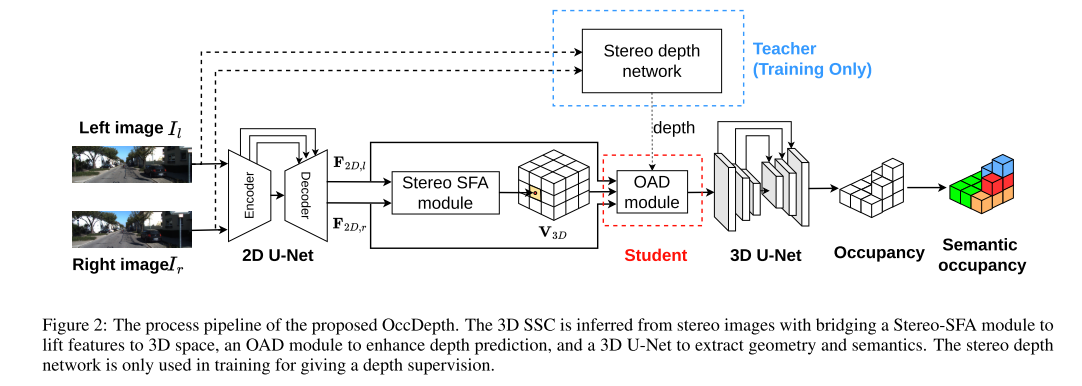
## 2.2 OccDepth: A Depth-aware Method for 3D Semantic Occupancy Network

OccDepth[3]的主要贡献：提出了一种基于双目图像预测三维语义栅格+语义补全的框架。

#### 2.2.1 算法整体流程

整体流程和MonoScene非常像（代码也是在MonoScene基础上改的）：

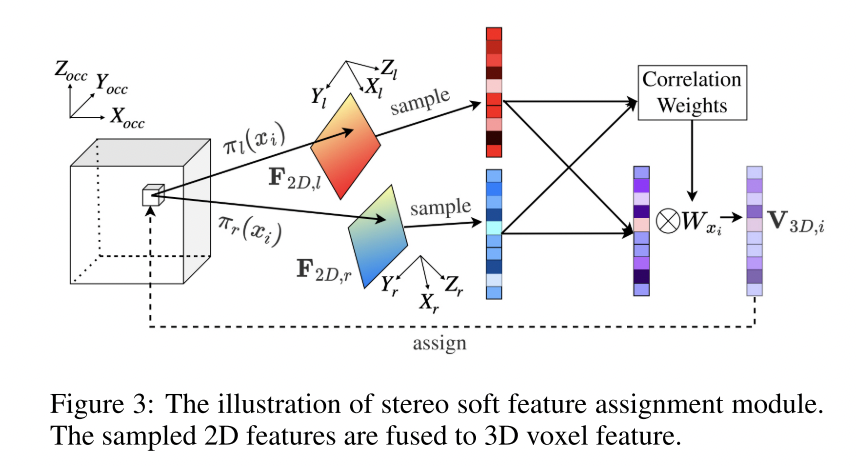
1. 2D UNet提取双目图像特征
2. 三维栅格从左右目图像中pull特征并融合
3. Occupancy Aware Depth Module利用深度估计对栅格特征进行类似attention的操作
4. 和MonoScene一样的UNet结构提取三维特征
5. 预测head输出每个栅格是否占据+语义类别



#### 2.2.2 图像特征提取及投影

**图像特征提取：**与MonoScene相同，采用预训练的EffecientNetB7。

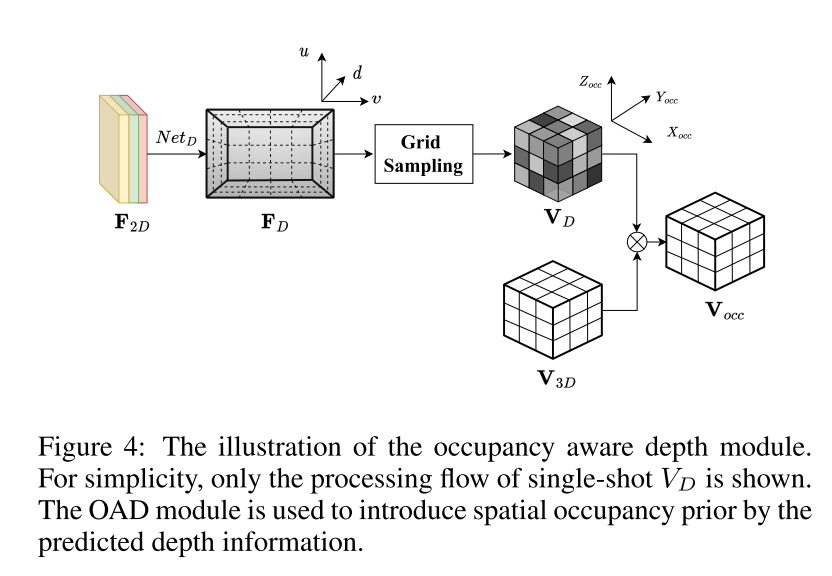
**特征投影：**左右目图像的特征投影与MonoScene相同，OccDepth的特殊之处在于：栅格的特征是将左右目的特征进行平均并乘以权重，权重的计算方法是左右目特征的Cosine Similarity；如果左右目没有重叠则使用单目的特征。个人感觉这个很有道理，因为这个计算特征相似度的过程相当于在进行双目匹配，如果左右目特征不吻合说明这个栅格很可能是空的。因此加上cos similarity应该可以帮助网络更好地学习双目匹配。



#### 2.2.3 Occupancy Aware Depth Module

除了Cosine Similarity，OccDepth还使用了单目深度估计来指导特征投影过程，OAD模块的主要流程是：

1. 左目/右目图像scale=8层的特征图经过一个网络提取Net\_D深度特征
2. 经过softmax，得到每个视锥对应的深度分布
3. 根据内参矩阵和外参矩阵，投影至栅格对应的坐标系，采样得到每个栅格对应的深度概率
4. 每个栅格对应的深度概率作为注意力权重，乘以上一步特征投影得到的特征。（对于有双目视野重叠的栅格，深度概率会取左右目的均值。）



#### 2.2.4 训练

Loss的计算和MonoScene基本一样，不同之处为：

* 为了帮助训练深度预测模型，作者使用双目图像估计的深度指导深度网络Net\_D的学习。具体是使用双目深度预测网络LEAStereo的深度预测离散化后的one-hot真值，用Cross-entropy Loss训练。
* 是否占据的预测和语义类别的loss分开计算。而MonoScene将是否free作为一个语义类别。

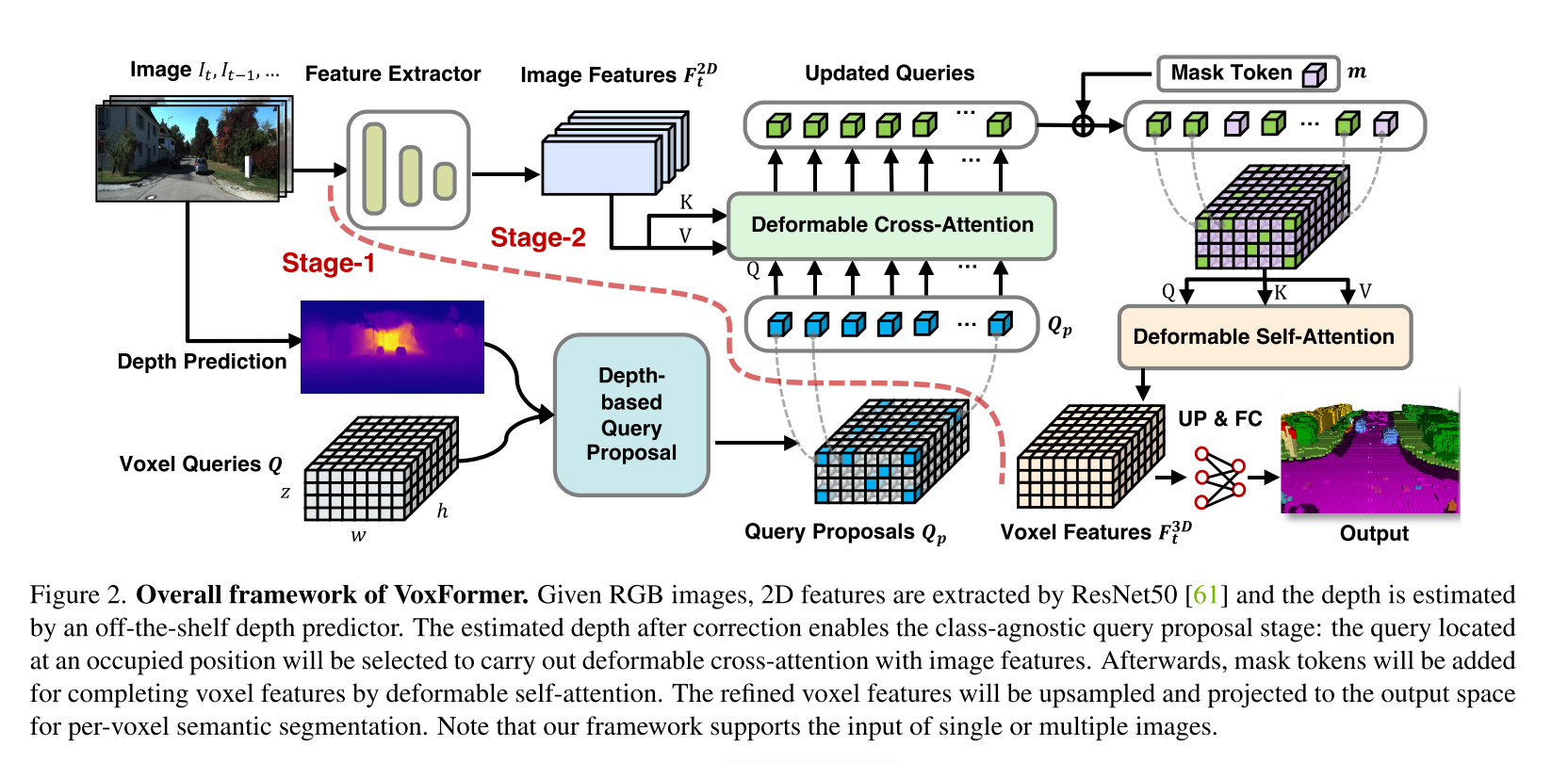
## 2.3 VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion，CVPR 2023

VoxFormer[4]的主要贡献：提出了一种基于单目图像序列预测三维语义栅格+语义补全的框架。

#### 2.3.1 算法整体流程

VoxFormer的整体流程有点类似Tesla Occupancy Network的方案，流程中大部分用到了Transformer：

1. ResNet-50提取图像特征
2. 生成一些类别无关的3D栅格Proposal用于后续query
3. 3D栅格Proposal作为query从图像中取特征（Cross Attention）
4. 类似MAE的流程，以3D栅格Proposal的特征为输入，通过Self Attention得到完整栅格特征
5. 预测head输出每个栅格是否占据+语义类别



#### 2.3.2 类别无关的3D栅格Proposal生成

这一步主要是选择一些在相机视野内的，比较重要的栅格。主要通过深度估计+占据概率预测的方式实现。具体流程如下：

1. **深度估计：**使用单目或双目深度估计网络预测每个像素的深度（实验中使用了MobileStereoNet），然后根据相机内参投影到相机坐标系中
2. **深度修正（占据概率预测）：**深度估计直接投影的点云存在很多噪声，因此使用基于3D栅格的occupancy预测网络。具体是将深度投影的点云进行0-1栅格化，然后用3D Unet架构输出栅格分辨率更低的占据预测。（实验中使用了类似LMSCNet的网络，LMSCNet是一个基于激光雷达的3D语义补全方法）
3. **query栅格生成：**上一步预测的所有占据栅格作为query栅格

#### 2.3.3 三维特征提取及输出

主要分为3个部分，query栅格和图像特征的Cross Attention，以及query栅格和其他所有栅格的Self Attention，以及最后的输出。其中attention均采用Deformable Attention（在Deformable DETR中提出的，BEVFormer也用到了）

**Deformable Transformer**我理解的流程是这样的： 给定Query按照某种方式得到一个参考点，同时经过几个可学习layer得到offset->参考点附近按照offset采样得到KV -> Transformer的Cross Attention。主要的优点是在Cross Attention时不需要使用所有的栅格的KV，而是通过可学习的方式选择一定数量的感兴趣的位置，从而减少计算量。具体细节可以参考大佬的这篇文章。

**Deformable Cross Attention** 主要是把每一个Query栅格的坐标投影到各帧图像上，投影点作为Deformable Attention的参考点（reference point），参考点附近采样一些点（可学习），这些点对应的图像特征作为Key，Value输入。同一个栅格可能会投到多帧图像上，那么多帧图像的attention输出进行平均，作为最终的输出。

**Deformable Self Attention** 是把所有栅格（HxWxZ）的特征作为输入，但对于非Query栅格，使用一个mask token替换原有特征。其中Deformable Attention参考点（reference point）为栅格中心坐标。

#### 2.3.4 训练

第一阶段的网络（类别无关的3D栅格Proposal生成）的loss为栅格是否占据的Cross-entropy Loss。

第二阶段的网络（Deformable Attention及输出层）的loss包括：

* 语义类别的Cross-entropy Loss
* MonoScene的Scene-Class Affinity Loss

## 2.4 TPVFormer: Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction，CVPR 2023

TPVFormer[5]的主要贡献：提出了一种三维特征的表示方法，通过三个正交平面上的特征来近似表示三维点/三维栅格的特征。这种表示形式可以用于3D点云语义分割、3D语义补全等任务。也是学术界对Tesla Occupancy Network的一种复现和改进方案。

#### 2.4.1 算法整体流程

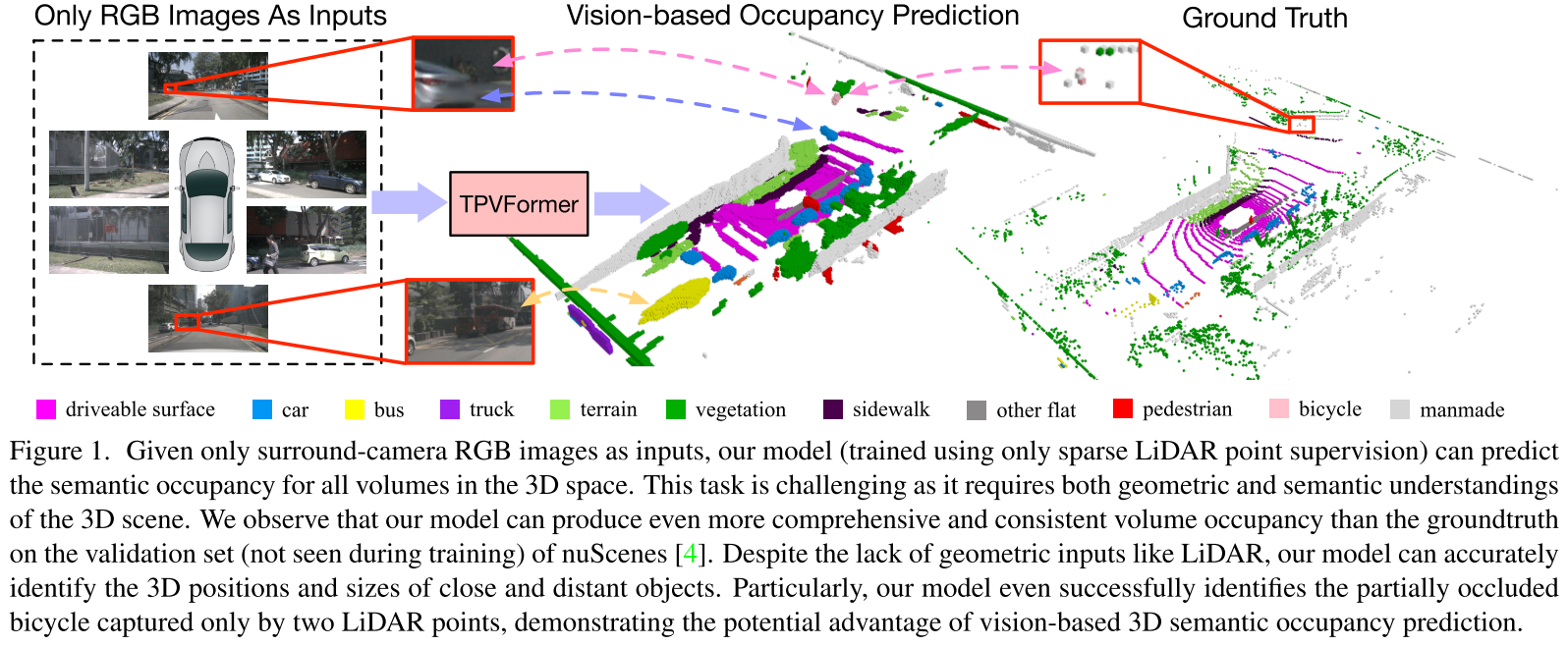
1. 根据任务类型定义一些点/栅格，比如语义补全任务可以初始化一些栅格

2. 每个栅格可以作为query，得到三个平面上的TPV query

3. 每个TPV平面的query经过若干层 ICA和CVHA层得到融合后的特征

4. 3个TPV特征加起来作为该栅格特征

5. 预测head输出该栅格是否占据+语义类别

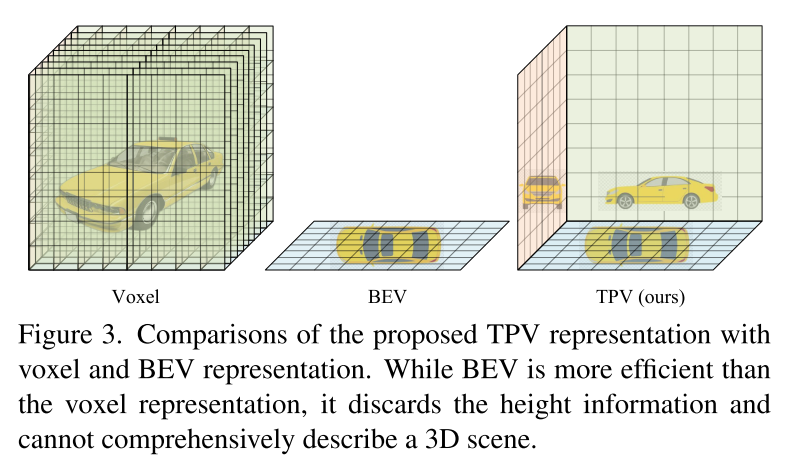


#### 2.4.2 TPV特征

作者先定义了TPV特征；然后给出了基于TPV平面上一个给定query点的两种Cross Attention，分别是和图像的attention以及和三个TPV之间的attention；最后基于这两种attention构建网络即为TPVFormer，完成各种感知任务。

由于三维栅格的空间和时间复杂度都是随着边长三次方增长，很难在真机实时部署；BEV虽然复杂度降低但高度方向的信息可能会有一定丢失，因此TPVFormer提出将空间的点向三个正交平面投影的方式（类似Convolutional Occupancy Networks中提出的2D版本），将三维点特征由三个平面的特征来表示。

TPV特征是在水平、径向、横向三个平面中分别保存的特征图。当下游任务需要获取三维点/三维栅格时，可将query的点/栅格投影到各平面，从各平面对应位置采样feature （双线性插值），然后通过某种aggregate方法（求和）得到三维特征。



#### 2.4.3 TPVFormer

TPVFormer包含两种attention，一种是TPV特征到图像的Cross Attention，另一种是三个特征平面之间的Cross Attention。

**TPV特征到图像的Cross Attention（ICA）:**和VoxFormer类似，采用Deformable Attention。对于一个TPV平面上（如XY平面）的某个像素（记为Query点），该点的垂直方向上采样一些 三维参考点（如不同的Z值），这些点投影到各个图像上，得到的二维坐标，作为Deformable Attention的参考点（记为二维参考点）。如果该点在所有图像视野外，则剔除该三维点。然后所有这些二维参考点附近采样一些点（可学习），作为Deformable Attention的KV。也就是说一个TPV平面上的Query对应N个图像的n\_z个参考点的附近的n\_offset个采样点的KV（因为有很多高度不同的三维参考点投影到各个图像，并在附近采样多个KV）,即一个Query对应N\*n\_z\*n\_offset个KV。同一个三维参考点从不同图像Cross Attention来的特征计算平均则为该三维参考点最终的的特征。

TPV特征平面间的Cross Attention（CVHA）

**和上一部分TPV到图像的Cross Attention类似：**对于一个TPV平面上（如XY平面）的某个像素（记为Query点），会在该点附近采样一些点，然后在这些点高度方向上采样一些点，投影到另外两个TPV平面上。

剩下的步骤和TPV到图像的Cross Attention一样了。

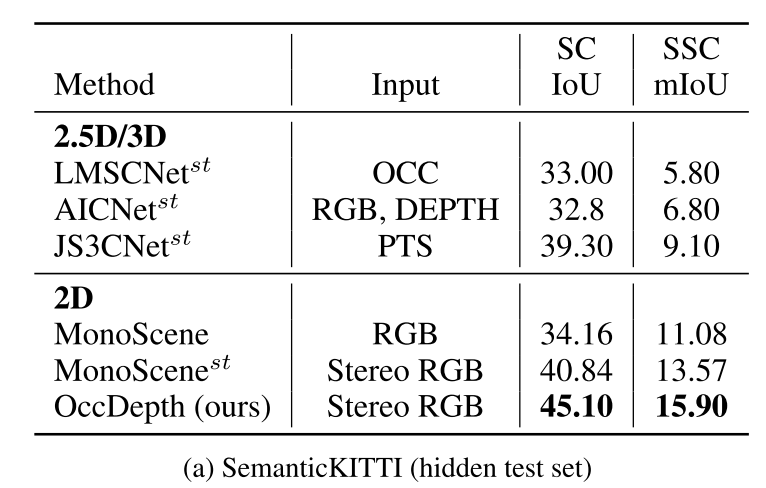
#### 2.4.4 训练

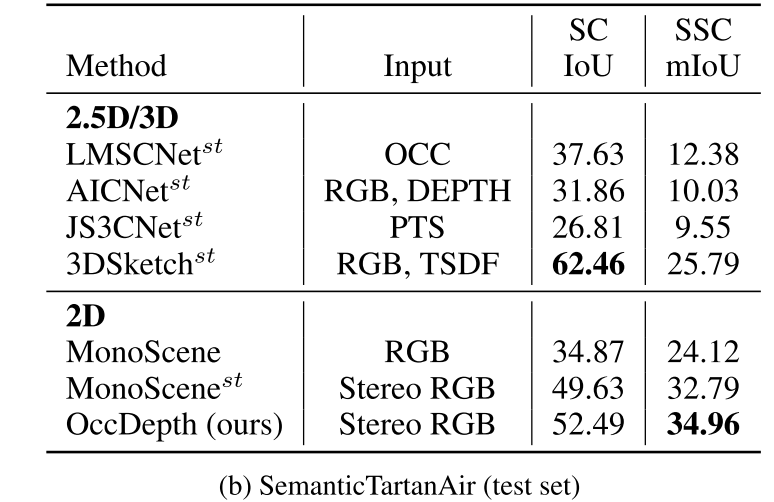
Loss的计算和MonoScene基本一样，不同之处为：没有使用relation loss 外的损失

# 研究工作

##### 3.1 occDepth

旷视研究院提出了OccDepth的⽅法，将纯图像输⼊⽅法的精度⼤幅提升，获得了视觉⽅法的SOTA，其中SC IOU从 34.2增⻓为 45.1，mIOU 从 11.1增长为15.9。笔者参与了该项工作的部分实验。



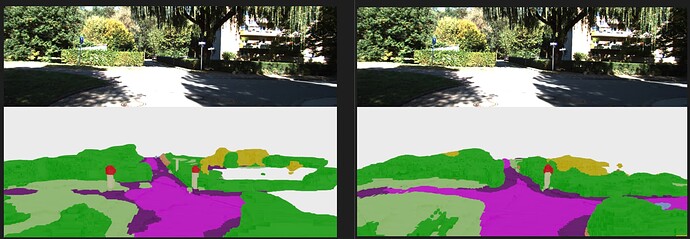


可视化对比图挑选。挑选原则：因为 LMSCNet、AICNet、JS3CNet 效果基本都远低于 OccDepth，所以挑选图片主要是挑 OccDepth 比 MonoScene 明显好的图片

###### 3.1.1 SemanticKITTI

（左 OccDepth；中MonoScene，右：GT）

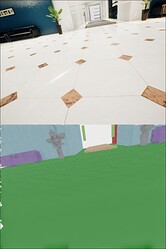
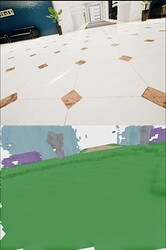
路标漏检（00080.png）

[](https://discourse.brainpp.cn/uploads/default/original/3X/1/0/10a4c50dab7bb8ebccb33a3f883801af01ccedaa.jpeg" \o "image)

###### 3.1.2 tartanair

（左 MonoScene；中 OccDepth； 右 gt）

绿植漏检+墙面空洞 (id: 50)

[](https://discourse.brainpp.cn/uploads/default/original/3X/7/9/79d621c3c32312f531ce8a98ebac6f08cac3ce51.jpeg" \o "image)

#### 3.2 CVPR2023-3D-Occupancy-Prediction

**三维占据栅格预测挑战赛**

在数据标注及训练过程中，3D边界框往往不足以描述一般物体，受机器人学概念的启发，可将感知表征描述成对栅格化三维空间的占据情况预测。这一想法在CES 2023和Tesla AI Day 2022上也得到印证。与以三维框为中心的感知表征不同，占据栅格享有描述任意形状实体的灵活性。

比赛任务：要求参赛者根据主办方提供的大规模三维占据栅格数据集，以及既定覆盖360度视角的多相机图像输入，预测出整个场景的栅格化三维空间中每个体素的语义状态。

•数据量：约30G

•参考训练时间：1-2天（8卡3090）

• Baseline（基础版）：miou=23.67 [https://github.com/CVPR2023-Occupancy-Prediction-Challenge/CVPR2023-Occ upancy-Prediction-Challenge](https://github.com/CVPR2023-Occupancy-Prediction-Challenge/CVPR2023-Occupancy-Prediction-Challenge)

• Baseline（进阶版，InternImage）：miou=25.11 <https://github.com/OpenGVLab/InternImage>

现在比赛未结束，不公开技术细节。

目前基于InternImage堆叠tricks的最佳点数是miou=43.11

基于VoxelFormer堆叠tricks的最佳点数是miou=44.5457

ensemble之后miou=46.8

# 参考文献

1. Shi, Yining, et al. "Grid-Centric Traffic Scenario Perception for Autonomous Driving: A Comprehensive Review." arXiv preprint arXiv:2303.01212 (2023).
2. Cao, Anh-Quan, and Raoul de Charette. "Monoscene: Monocular 3d semantic scene completion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
3. Miao, Ruihang, et al. "Occdepth: A depth-aware method for 3d semantic scene completion." arXiv preprint arXiv:2302.13540 (2023).

[4] Li, Zhuoling, et al. "VoxelFormer: Bird's-Eye-View Feature Generation based on Dual-view Attention for Multi-view 3D Object Detection." arXiv preprint arXiv:2304.01054 (2023).

[5] Huang, Yuanhui, et al. "Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction." arXiv preprint arXiv:2302.07817 (2023).