BEV检测

目录

自底向上BEV特征建模

可以把"自底向上"理解为由"2D to 3D"借助深度估计或者3D编码，先把图像提升到点云(2d to 3d)，进行voxel pooling成BEV.而"自顶向下"理解为"3D to 2D"，先生成BEV query，再用query投影到图像(3d to 2d)，对图像特征attention来query出3D特征

Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D

https://github.com/nv-tlabs/lift-splat-shoot

LSS是早期的比较直接的尝试，即先估计每个像素的深度，再通过内外参投影到bev空间。只是因为不存在深度标签，这里并没有直接回归深度值，而是对每个像素点预测一系列的离散深度值的概率，概率最大的深度值即为估计结果。

可以得到深度分布特征α和图像特征c，将二者做外积，可以得到一个视锥特征（frustum-shaped point cloud）

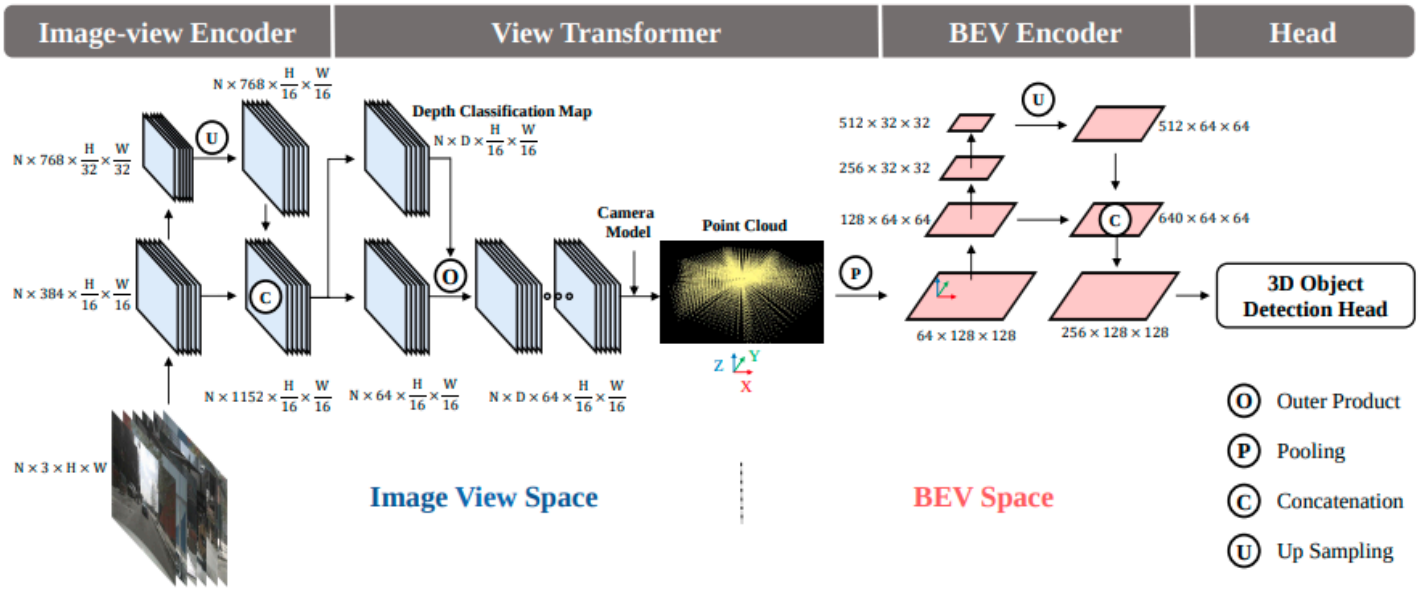
得到多视角的视锥特征后，可以通过外参将视锥投影到bev平面。在bev平面下，每个存在高度信息的像素称为体素（voxel），具有无限高度的voxel称为pillar，将每个视锥的每个点分配给最近的pillar，再执行sum pooling，得到CxHxW的bev特征。作者采用cumsum trick来提升sum pooling 效率，并把这一过程称为splat.

有了bev特征后，就可以很方便的进行3D检测、语义分割、预测和规划等一系列任务，作者把这个过程称为shoot。LSS方法可以得到稠密的bev特征，缺点是由于每个像素都预测了一系列深度概率值，计算量相对较大。

BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View

基于LSS的自底向上建立BEV的方法

先对多视角图像进行特征提取,再通过基于LSS的视角转换（View Transformer）将多视角特征投影到bev空间下，再用和第一步类似的backbone对bev特征进行编码，最后进行目标检测。这种方法虽然在LSS这一步存在不少冗余的计算，但好处是得到了显式的bev特征，可以做bev视角下的特征提取和数据增强，并且可以使用任意的目标检测头。



提出了scale-NMS，即对不同类别的目标进行不同尺度的缩放，来做更符合客观场景的目标框过滤

自顶向下bev 3D检测

DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries

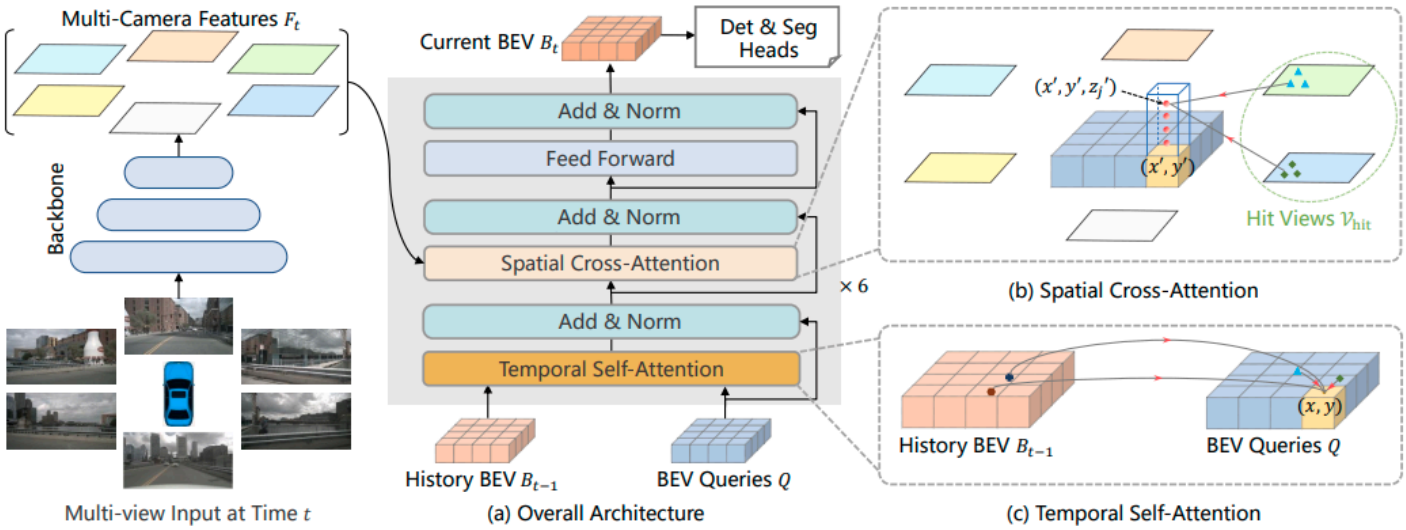
https://github.com/WangYueFt/detr3d

由于bev特征需要从多视角图像特征融合得到，用Resnet+FPN（没有transformer encoder模块）对多视角图像提取特征。Decoder模块参照deformable DETR的思路，在bev空间预设多个3D的object queries，并从object queries经线性映射得到3D的参考点（reference points)。下一步是3D的参考点如何与2D的特征做交互，文中利用了内外参的先验信息，将3D reference points投影到各个视角的图片上。由于多相机之间存在共视区域和盲区问题，一个参考点可能投影到多个视角，也可能一个视角也投不到，所以加了一个二进制的mask代表当前视角是否被投影成功。

接下来是做cross-attention，DETR3D的做法与DETR和deformable DETR都有一些不同，object queries不是和DETR那样与全图交互，也不是和deformable DETR那样先从object queries预测一些参考点，再预测一些以参考点为基准的采样点，然后和采样点的特征交互，而是直接和3D参考点投影的2D参考点处的特征交互（经过双线性插值），相当于交互的特征个数=object queries个数，比deformable DETR还要少（每个object query预测K个采样点，默认是4个），应该说是更稀疏的deformable DETR了。后面bbox推理值和真值的匹配和损失函数的计算和DETR是一样的。

BEVFormer

采用纯视觉



整体pipline：

- Backbone + Neck （ResNet-101-DCN + FPN）提取环视图像的多尺度特征；
- 论文提出的 Encoder 模块（包括 Temporal Self-Attention 模块和 Spatial Cross-Attention 模块）完成环视图像特征向 BEV 特征的建模；
- 类似 Deformable DETR 的 Decoder 模块完成 3D 目标检测的分类和定位任务；
- 正负样本的定义（采用 Transformer 中常用的匈牙利匹配算法，Focal Loss + L1 Loss 的总损失和最小）；
- 损失的计算（Focal Loss 分类损失 + L1 Loss 回归损失）；
- 反向传播，更新网络模型参数；

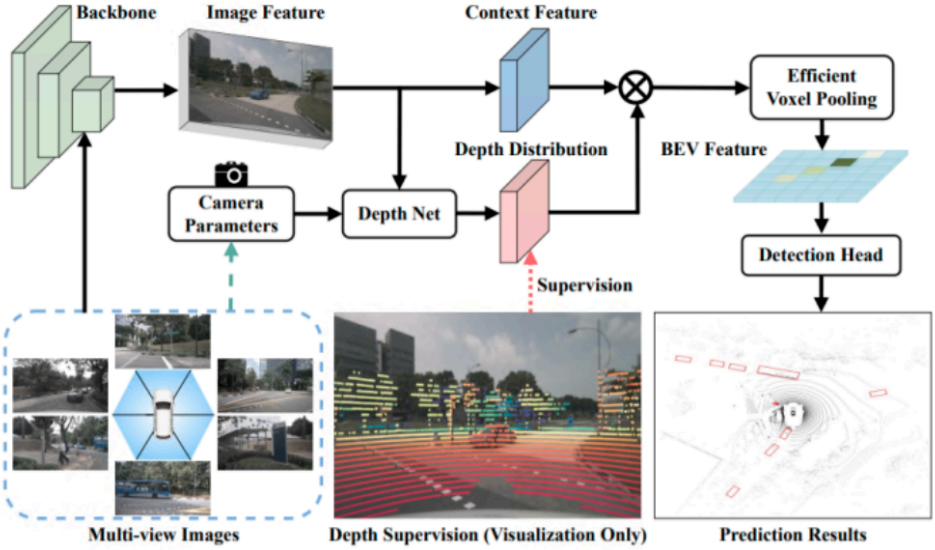
输入的数据是一个 6 维的张量：（bs，queue，cam，C，H，W）

bs：batch size；queue：连续帧的个数；cam：每帧中包含的图像数量，对于nuScenes数据集而言是六张环视图片；

C，H，W：图片的通道数，图片的高度，图片的宽度；

SCA: Spatial cross-attention

BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection



BEVHeight++: Toward Robust Visual Centric 3D Object Detection

TABLE 3: **Comparison on the nuScenes *val* set.** "L" denotes LiDAR, "C" denotes camera and "D" denotes Depth/LiDAR supervision. † denotes initialization from an FCOS3D [53] backbone.

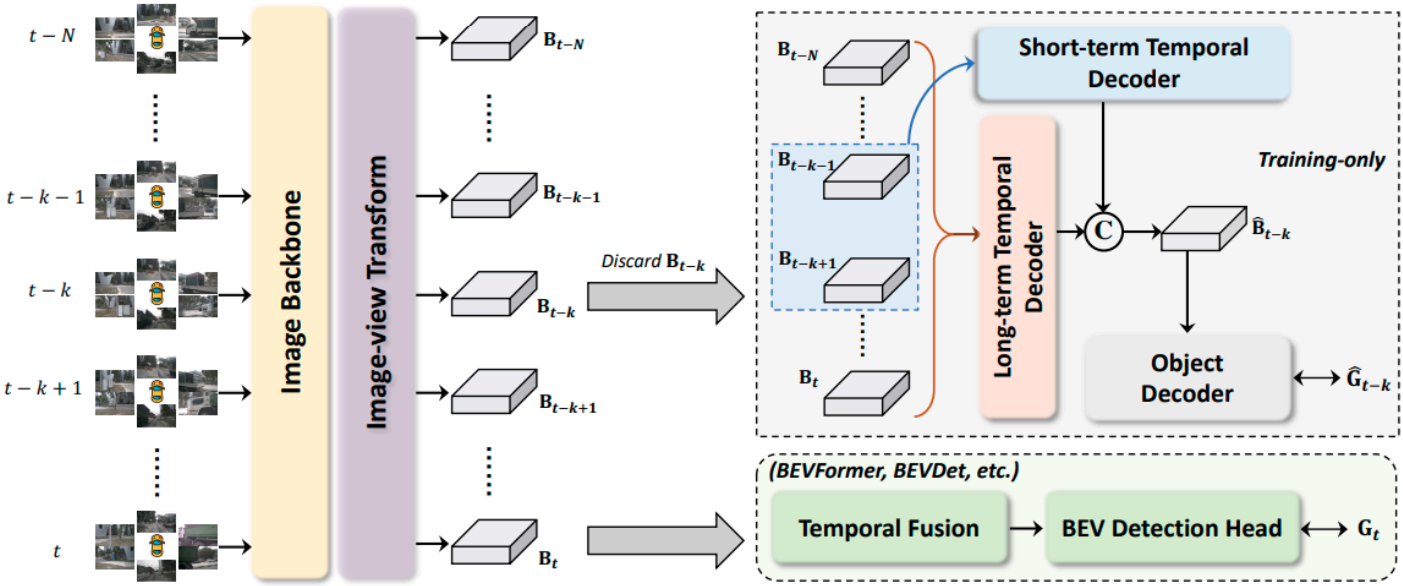
Methods	Backbone	Image Size	Modality	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterPoint-Voxel [54]		-	L	0.648	0.564	-	-	-	-	-
CenterPoint-Pillar [54]		-	L	0.602	0.503	-	-	-	-	-
FCOS3D [53]	R101-DCN	900×1600	C	0.415	0.343	0.725	0.263	0.422	1.292	**0.153**
DETR3D† [20]	R101-DCN	900×1600	C	0.422	0.347	0.765	0.267	0.392	0.876	0.211
DETR4D† [55]	R101-DCN	640×1600	C	0.509	0.422	0.688	0.269	0.388	0.496	0.184
PETR† [21]	R101-DCN	900×1600	C	0.442	0.370	0.711	0.267	0.383	0.865	0.201
PETRv2† [22]	R101-DCN	640×1600	C	0.524	0.421	0.681	0.267	0.357	0.377	0.186
PolarFormer† [26]	R101-DCN	900×1600	C	0.528	0.432	0.648	0.270	0.348	0.409	0.201
BEVFormer† [25]	R101-DCN	900×1600	C	0.517	0.416	0.673	0.274	0.372	0.394	0.198
BEVDet [29]	Swin-T	512×1408	C	0.417	0.349	0.637	0.269	0.490	0.914	0.268
BEVDet4D [32]	Swin-T	640×1600	C	0.515	0.396	0.619	0.260	0.361	0.399	0.189
Fast-BEV [56]	R101	900×1600	C	0.535	0.413	0.584	0.279	0.311	0.329	0.206
SOLOFusion [57]	R101	512×1408	C	0.544	**0.472**	**0.518**	0.275	0.604	0.310	0.210
BEVDepth [7]	R50	256×704	C&D	0.475	0.351	0.639	0.267	0.479	0.428	0.198
BEVHeight++	R50	256×704	C&D	0.498	0.373	0.614	0.269	0.419	0.375	0.203
BEVDepth [7]	R101	512×1408	C&D	0.535	0.412	0.565	0.266	0.358	0.331	0.190
BEVHeight++	R101	512×1408	C&D	**0.554**	0.423	0.541	**0.262**	**0.307**	**0.277**	0.187

BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection

BEVerse: Unified Perception and Prediction in Birds-Eye-View for Vision-Centric Autonomous Driving

历史帧物体预测（HoP）Temporal Enhanced Training of Multi-view 3D Object Detector via Historical

Object Prediction



首先，利用图片的backbone和视角转换网络得到从t到t-N时刻的BEV特征，并丢弃第t-k帧的BEV特征信息。

其次，设计了一种时间解码器，用来在剩余帧的BEV特征中提取有价值的信息，重建一个虚拟的t-k帧的BEV特征。

该时间解码器包括长期时序信息捕捉分支和短期时序信息捕捉分支。短期时序信息捕捉分支重点在于提取空间语义信息，主要利用t-k帧前后两帧；另一方面，长期时序信息捕捉分支则能够提取物体的运动信息，利用的是其余所有帧的信息。

在虚拟的t-k帧的BEV特征上，增加了一个轻量的BEV检测头来预测t-k帧的物体。

除了HoP，论文中还提出了历史帧Query融合（Historical Temporal Query Fusion），可以从Query层面融合历史帧的信息来帮助当前帧的检测。