1. **Motivation & Contribution**

动机：

1. 当前环视场景的感知，大多基于单帧单目检测网络，通过后处理将最终的结果进行融合。该方案无法进行跨相机的信息融合，精度和效率较低。
2. 当前已有的BEV方案，需要首先预测离散的深度，如果深度预测不准，则导致BEV feature的误差较大，影响最终的检测结果**。BEVFormer则提出了一种不依赖depth预测，生成BEV feature的方法。**
3. BEV方案中显示的生成BEV feature对于时空的融合具有天然的优势。

贡献：

1. 提出了一种时空的Transformer结构，能够实现多相机以及时序上的融合。
2. 设计了BEV queries，通过上述的Transformer结构，查询空间与时序特征，可以得到最终的BEV feature。

**2. Method**

**2.1 框架**

**网络输入：**

1. 环视6张相机图像。
2. 前一帧BEV feature

**网络结构：**

1. 用于提取图像特征的backbone
2. 基于Transformer的时空编码器
3. 用于查询并生成BEV feature的BEV queries
4. 与Deformable DETR类型的检测头

**网络前向：**

1. 环视6张相机图像通过backbone，得到各自的image feature
2. BEV queries与前一帧BEV feature经过Temporal self-attention得到新的BEV queries
3. 新的BEV queries与6相机的特征做spatial cross-attention，并经过feed forward得到最终的BEV feature
4. BEV feature经过与DETR类似的decoder head得到最终的检测结果

**2.2 Deformable Attention Module**

**BEVFormer中的attention模块使用了Deformable Attention模块**

**输入：**query  zq，query的ref. point pq，以及context feature  x

**前向：**

1. query zq分别经过一个Linear层，生成相对于ref. point的偏差和特征聚合使用的权重
2. 利用ref. point以及offset，通过插值的方式在feature x获取对应的特征
3. 使用生成的权重对特征进行加权求和

**2.3 Temporal Self-Attention**

Q: BEV queries

Qp: 位于 p位置的query

p: BEV grid上的位置

V: context feature

Bt−1: 前一帧的bev feature

B′t−1: 利用ego-motion将前一帧对齐到当前帧之后的bev feature

BEV Querise不仅会与自身做self-attention，还会与前一帧的feature做cross-attention

与deformable attention不同，这里offset是使用 Q与 B′t−1concat之后的特征进行预测

**2.4 Spatial Cross-Attention**

Ft: t时刻6相机的图像特征

Fit: t时刻第i个相机的图像特征

Nref: 参考点数目

νhit: 投影点所在相机视角

P(p,i,j): 将参考点投影到图像的投影函数

**前向：**

1. 首先将bev grid在高度方向采样 Nref个点，高度方向-5米到3米， Nref=4
2. 有了bev坐标的x, y, z，通过内外参将该点投影到6个相机图像上，得到ref. points
3. query Qp分别与投影点所在的相机特征做cross-attention

**3. Experiment**

**实验设置：**

**backbone：**ResNet101-DCN与VoVNet-99

**bev grid设置：**

nusc: 200 x 200， [-51.2m, 51.2m]

waymo: 300 x 220， [-35m, 75m]

**BEVFormer-S指不使用前一帧的BEV feature**

**3.1 与SOTA的比较**

**3.2 多任务实验**

多任务训练相较于单任务，在精度上并没有落后

**3.3 Spatial Cross-Attention实验**

使用deformable attention的local相较于全局的attention精度有提高，同时更节省显存。

**3.4 Temporal Self-Attention实验**

1. 使用时序信息能够有效提升速度、位置以及朝向的精度
2. 使用时序信息能够提高遮挡目标的召回率

**3.5 时延与精度评估**

**4. BEVFormer V2**

| **BEVFormer  v2** | **BEVFormer v2: Adapting Modern Image Backbones to Bird’s-Eye-View Recognition via Perspective Supervision** |
| --- | --- |
| paper | [[2211.10439] BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision (arxiv.org)](https://arxiv.org/abs/2211.10439) |

**4.1 V2相较于V1的主要改进**

1. 在image backbone的基础上，添加了一个perspective 3d head，与bev head是并行关系。对backbone而言，相当于添加了一个额外的监督，使backbone更快的学习3D任务
2. 在encoder中，调换了spatial与temporal的位置，先进行spatial encoder再进行temporal encoder
3. 改进了temporal encoder，使用了前多帧的BEV feature，而V1只使用了前一帧
4. 设计了一种两阶段方式，使用PV 3D head的检测结果，作为BEV head decoder中的ref. point

**4.2 PV Loss**

**PV 3D head使用FCOS3D head**

**4.3 Ravamped Temporal Encoder**

1. 首先获取不同时刻的bev feature  Bk
2. 将不同时刻的 Bk对齐到t时刻 Btk
3. 将bev feature在channel维度上concat，并使用residual block进行降维
4. 降维之后的feature输入到temporal encoder

**4.4 Two-stage BEV Detector**

BEV head使用Deformable DETR Decoder layer

主要包含以下几个部分：

1. 输入的BEV feature
2. positional embeddings用于生成ref. point, 一个query对应一个positional embedding
3. query用于查询目标，一个query对应一个位置中的目标
4. 使用PV head检测出的结果，经过后处理过滤并投影到bev 平面上得到额外的ref. point，辅助训练。此时一个query怎么与ref. point匹配？

**4.5 实验**

**4.5.1 与SOTA的比较**

InternImage-B与V2-99参数量大致相同

InternImage-B是在COCO数据上经过2D检测任务预训练，而V2-99是经过depth预训练的，并在nusc上fine-tuning之后的

**4.5.2 PV监督消融实验**

BEV优于PV，PV+BEV的两阶段精度更优

**4.5.3 采用不同backbone的效果**

对于不同的backbone都能有效提升精度

**4.5.4 使用PV监督对于收敛速度的影响**

使用PV监督在48epoch时即可收敛，同时精度更高

**4.5.5 不同head的影响**

DD3D head相对于DETR3D head是一种dense的预测，对于backbone更有利

group detr优于group detr

**4.5.5 不同trick的影响**

1. image-level数据增强：图像的水平翻转
2. 更长的时间间隔，由v1的0.5s增大到2s
3. 双向temporal encoder，使用后几帧的bev feature