多模态融合学习笔记

模态融合的种类

图像之间融合：1.不同曝光程度的图像之间进行融合[5,6,13]；2.医疗图像之间融合(PET & CT)[11]；3.通用型图像融合框架[8]

图像和自然语言融合：1.自然语言引导的图像分割[7]

图像和探测器信号融合：1.图像和3D点云融合[1,4,9,10,12,14]；2.图像和热成像图像融合[2,3]

模态融合的方式

提取特征后并转换到同一特征空间后，利用cross-attention等机制进行特征融合[4,7,12,14]

在encoding过程中逐步融合[2,3]

分别进行encoding特征提取后，融合特征图[5,10,13]

在图像编码前融合(RGB图像转换为YCbCr并在Y channel进行融合)[6,8]

根据不同模态生成对应的策略，进行策略融合[9]

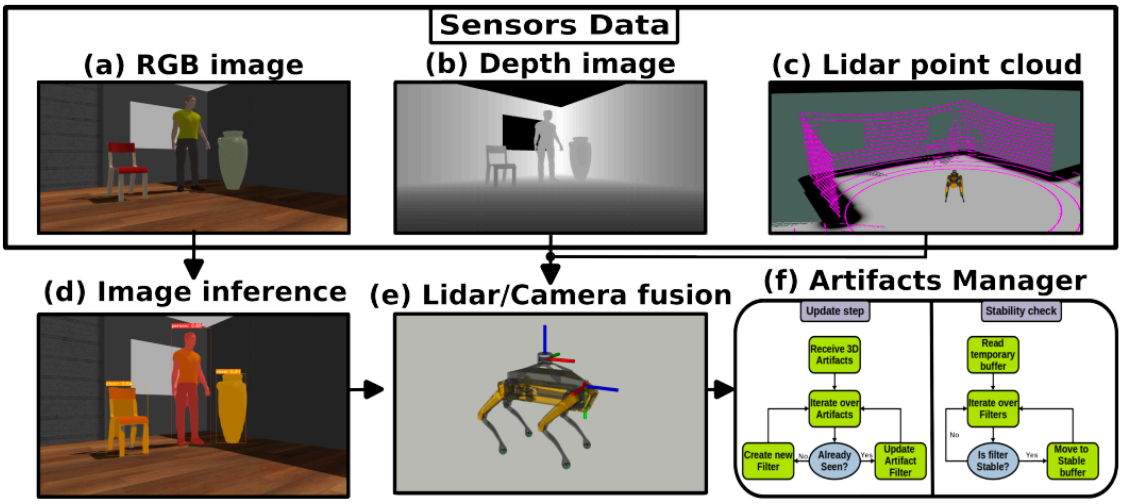
分别编码两种模态，并依据结果进行对比学习[11]

各文章简述

[1]Artifacts Mapping: Multi-Modal Semantic Mapping for Object Detection and 3D Localization

这篇文章的应用场景是运用在机器人上的场景理解。在该工作中，算法维护一个预设好的地图并在地图上利用lidar和RGB-D数据完成对物体的识别和定位。在这个框架中，算法首先利用RGB图像理解环境信息，之后利用多模态传感器信息进行特征融合提取depth信息，最终处理artifacts。

该工作的核心是一个online的多传感器信息融合语意映射框架。在感知模块，算法先利用2d图像进行物体分割，之后利用摄像/雷达信息进行3d物体位置估计。其中，在3d物体位置估计部分，根据3d点估计和相机的距离，改变相机和雷达间的权重。利用过滤后的摄像和3d点云信息，实现粗略的对于物体的3维半径的估计。



为了实现artifacts管理，算法分别进行位置过滤和位4置稳定。算法维护一个artifacts的临时buffer，当接收到artifacts时，算法检查该artifacts是否在buffer中存在，如果存在则更新buffer，如果不存在则在buffer中创建一个新实例。在位置稳定阶段，算法维护一个stable buffer，如果artifacts可以保持稳定，则会被移入stable buffer中。最后，属于同一个类的artifacts会重叠融合。

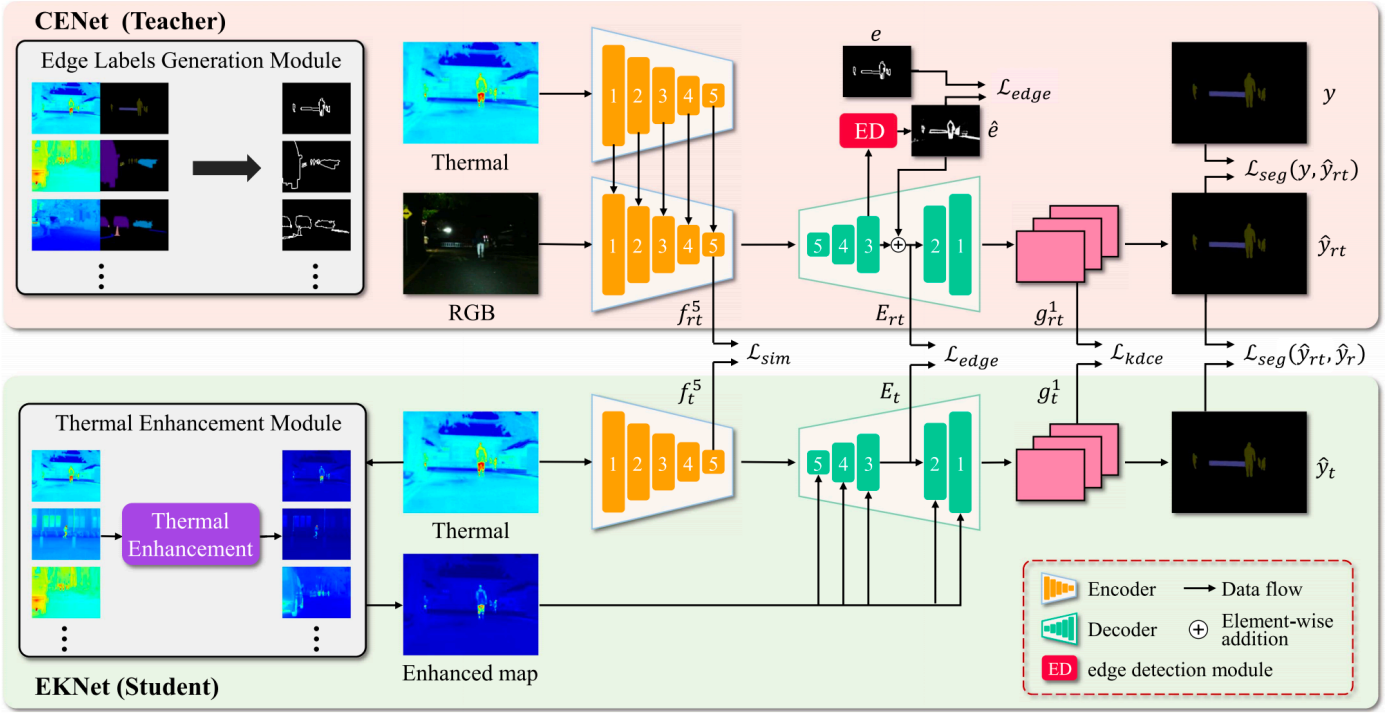
在实验部分，作者测试了该算法是否能够正确的探测和定位物体。针对不同传感器，作者测试了 只使用相机/只使用lidar/混合使用 带来的误差，证明了多模态融合方法的优越性。作者在实际场景中测试了相机和lidar生成的点云，验证了相机更适合近距离探测而lidar则适合远距离探测，这同样证明了多模态混合使用的必要性。

总结：该工作利用lidar和rgb-d摄影信息获取高精度的近距离和远距离物体信息（只用相机会丢失远距离信息，只用lidar会缺失高层次环境纹理信息）。同时作者开发了一个ui界面，可以在mapping过程中和artifacts map交互。

提升方向：针对运动的物体进识别和跟踪；

[2]CEKD:Cross-Modal Edge-Privileged Knowledge Distillation for Semantic Scene Understanding Using Only Thermal Images

这篇文章的应用场景是利用热成像进行的图像语意信息理解。目前已有的方法分为三种：只利用RGB图像特征的方法在照明条件不足的情况下性能显著下降，只利用热成像图像的方法在丢失了色彩信息且导致了边界模糊，同时利用这两种模态的方法需要高精度的传感器调制，对于有震动干扰的车辆搭载环境适用性不强。



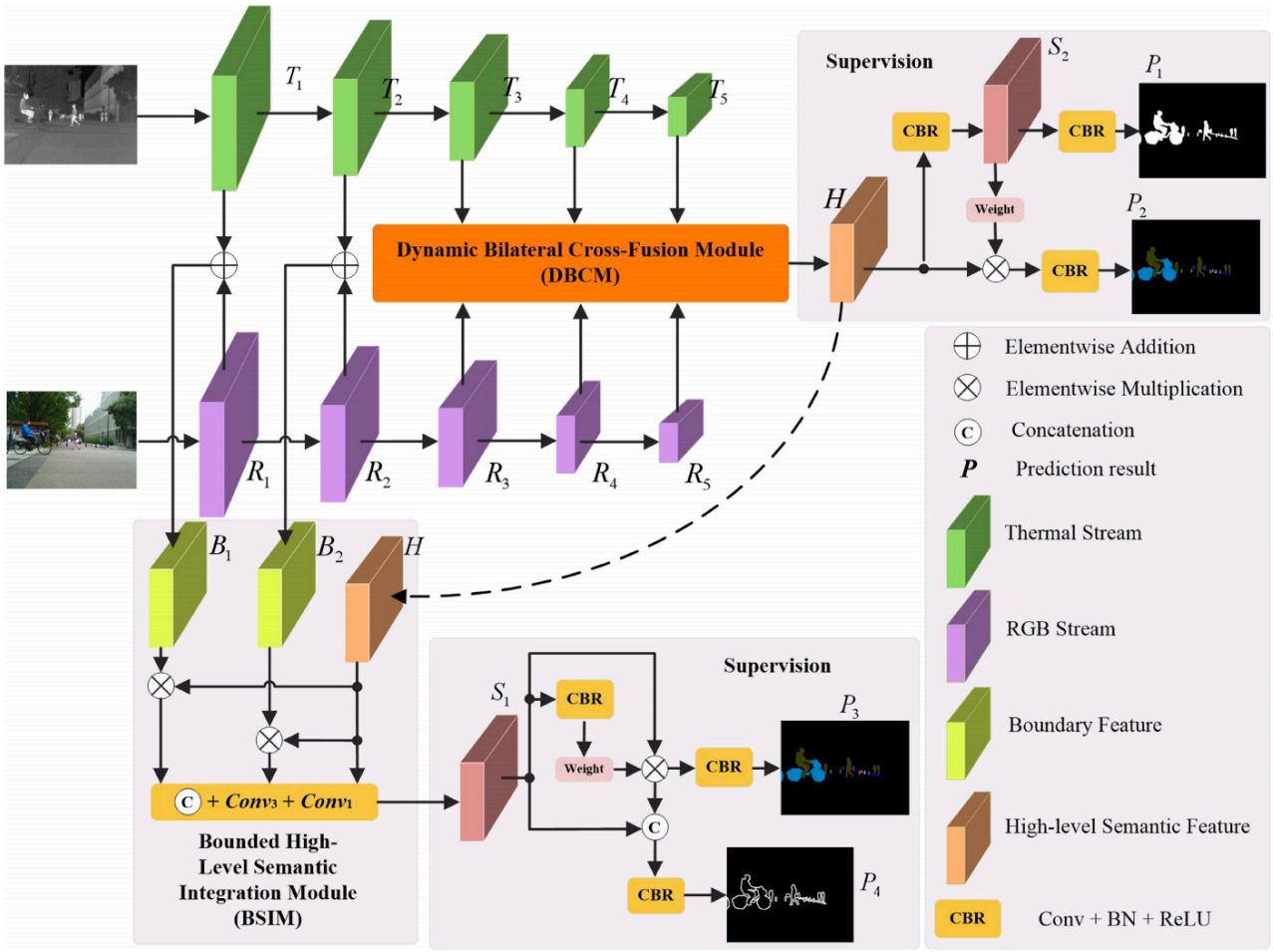
这篇文章提出了是利用双模态训练一个teacher 网络，并利用热成像图像进行知识蒸馏训练一个student网络，实现只依赖单模态进行语意理解的同时不丢失色彩和边界信息。

在实验部分，作者针对算法的不同部分进行了消融实验，对比了不同版本算法边缘检测的能力。之后将算法和其它sota方法进行了对比。

提升方向：1.RGB增强模块丢失了低对比度(温差)信息；2.由于student网络的encoder输入数据较少，想要拟合出和teacher网络相似的编码结果，需要改变设计，例如使用更多输入信息/更深的神经网络。

[3]DBCNet:Dynamic Bilateral Cross-Fusion Network for RGB-T Urban Scene Understanding in Intelligent Vehicles

这篇文章的应用场景是应用于自动驾驶的城市场景理解。文章中提出了利用RGB图像信息和热成像图像信息的多模态方法，这样的框架可以高效的实现多层次的特征融合。文章提出了DBCM模块，进行多层次的特征融合并充分提取高层次的语意信息。文章提出了BSIM，实现高层次语意信息和边界线索的融合，增强对城市场景的理解。



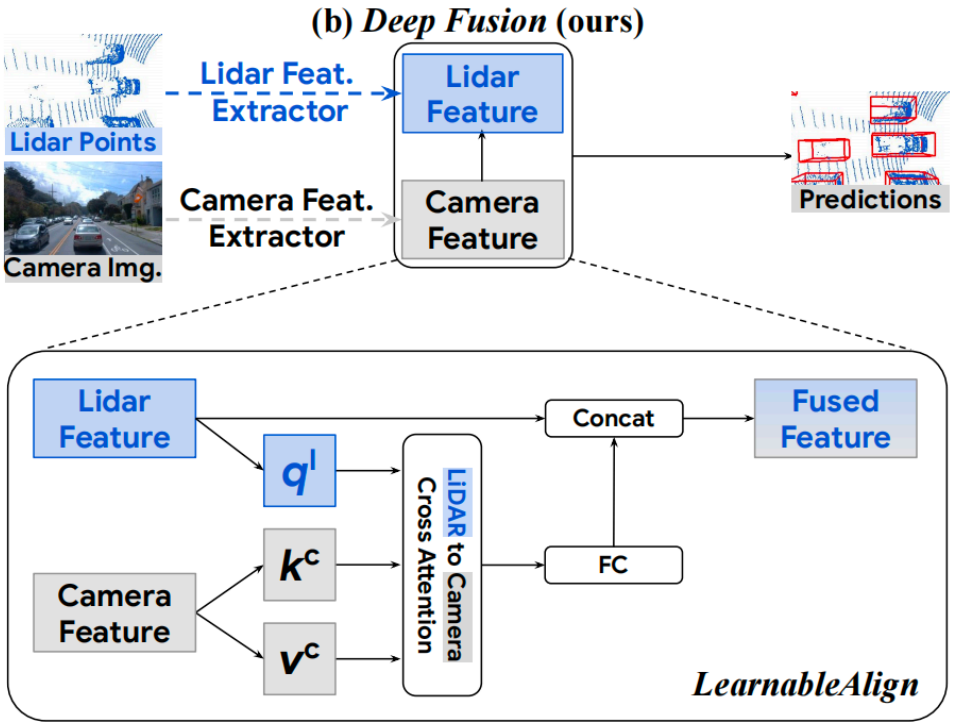
在两个模态的编码过程中，逐步进行特征融合。其中，信息增强模块(IEM)利用多种扩张卷积拼接的方式获取更多信息。由于在encoding过程中，各步骤的feature map 大小不一致，作者提出了动态交叉融合模块对特征进行融合。

在实验部分该方法与其它sota方法进行了对比，并且对各部分进行了消融实验。

提升方向：根据作者表述，该方法对于复杂图像和小目标的分割问题表现不佳。

[4]DeepFusion:Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection

这篇文章针对已有方法在将激光雷达信息和2维摄像信息进行融合使用方面的不足进行改进，提出将摄像特征和深度雷达特征融合的方法（而非直接的点云特征）。同时，由于这样的深度特征已经被高度的强化和合成，需要有一种在这样情况下辅助将点云信息和摄像信息一一对应的方法。



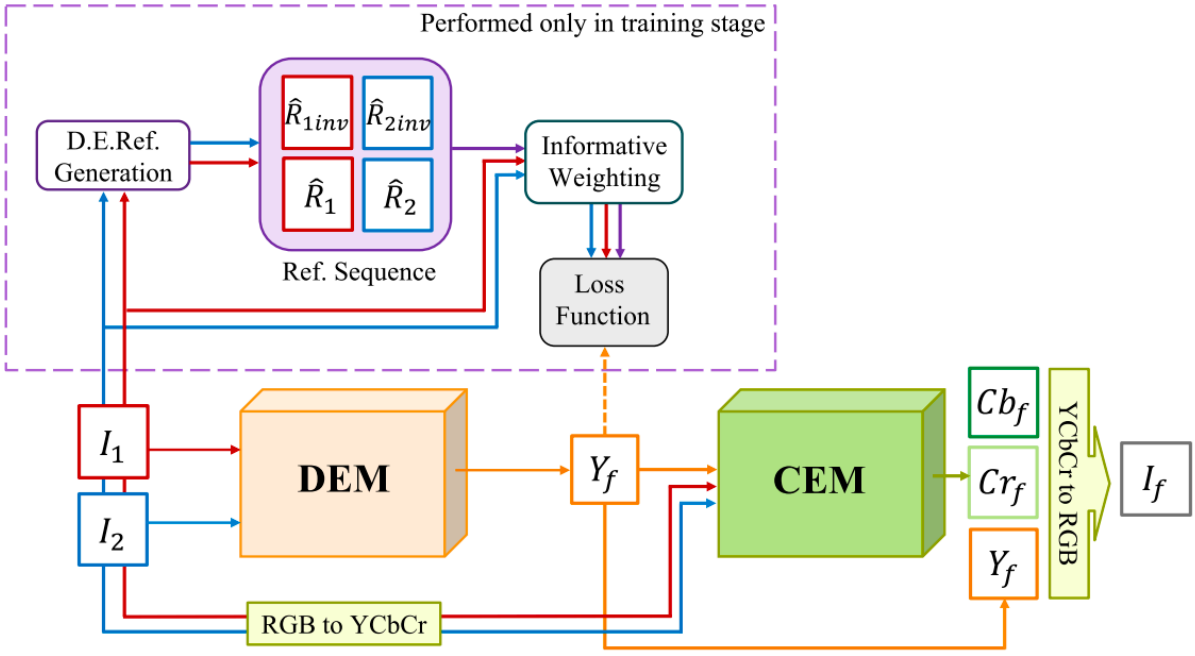
为了解决这些问题，文章首先提出了**InverseAug** 反转数据增强过程，并建立点云信息和2维摄像信息的一一对应。同时，在**LearnableAlign**部分利用cross-attention学习雷达信息特征和对应的2维影像特征。

在实验部分，作者和其它sota方法进行了对比，并研究了**InverseAug** 和**LearnableAlign** 的效果。

提升方向：1.反转过程只能针对特定的物理系统，不同的物理系统可能需要训练不同的算法。2.该方法对硬件系统的calibration要求较高。

[5]Multi-exposure image fusion via deep perceptual enhancement

这篇文章的应用场景是多曝光度图像融合问题。文章提出需要同时考虑重建图像的信息和视觉真实性，基于此，文章提出了一个深度感知网络。针对图像融合的机制和视觉质量，模型包括了细节强化和色彩强化模块。



通过引入细节增强模块DEM和色彩增强模块CEM，实现对于源图像信息的充分获取和重建图像的高视觉效果。

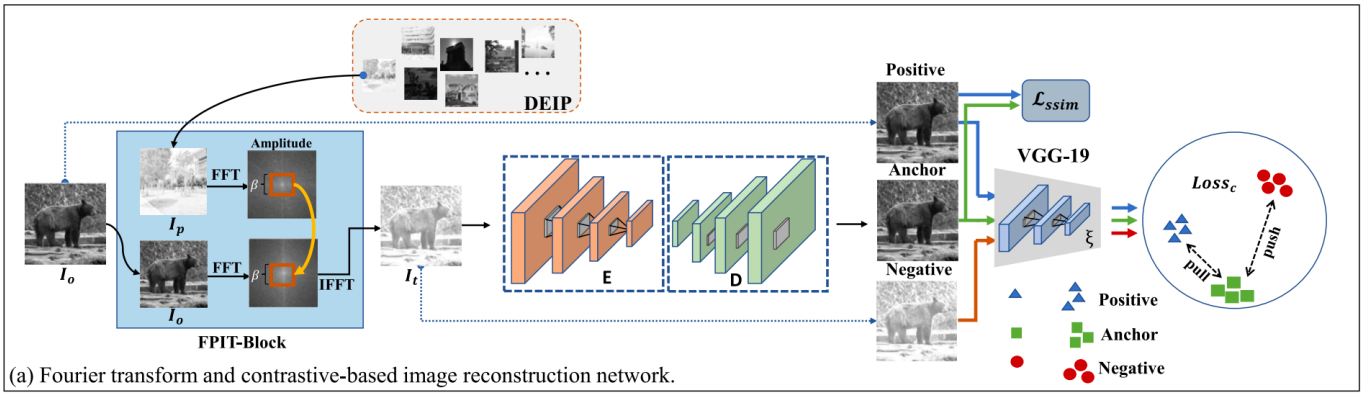
为了生成参考图像，作者对源参考图像进行了反转，实现双向的细节增强，每张参考图像对应两张增强后的图像。

在实验部分，作者定量&定性分析了模型的表现，并和其它sota方法进行了对比。针对损失函数和不同的参考图像组合进行了消融实验。

提升方向：1.现有方法只适用于静态图片，而不适用于动态场景；2.该方法目前无法调整融合的图片数量。

[6]Rethinking multi-exposure image fusion with extreme and diverse exposure levels: A robust framework based on Fourier transform and contrastive learning

这篇文章的应用场景是多曝光度的图片的融合问题。作者认为现有方法对于极端和多样化曝光度图像的融合问题不robust。文章提出了一个基于傅立叶变换和对比学习的MEF网络结构(FCMEF)。



文章利用自然图像数据集和具有多种曝光度图像的数据集构成网络的数据集，克服数据集不足的问题。利用基于傅立叶变换的像素强度转换策略实现在保留图像内容的同时改变其曝光度。

文章提出了一个对比正则损失(CR-Loss)增强融合网络的重建能力。

文章构建了两个MEF benchmark test sets。eMEFB(extreme exposure level) ,rMEFB(random image pairs)

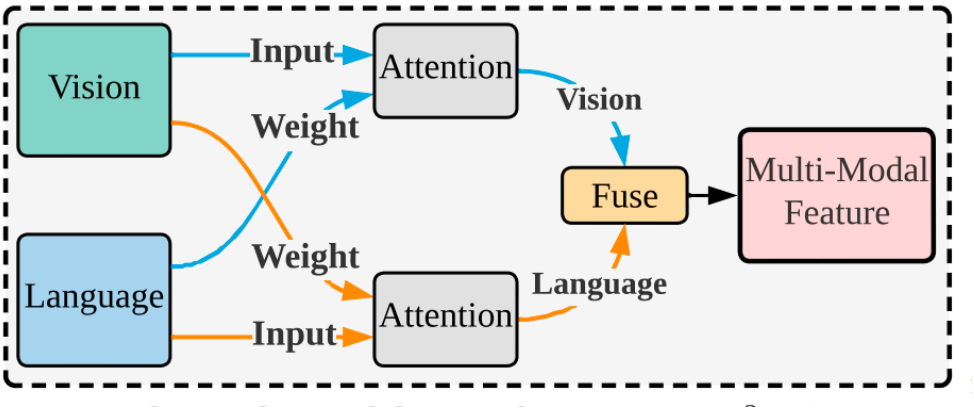
在实验部分，作者在一个已有的测试集和文章提出的两个测试集上进行实验。在基于人类感知的指标上进行对比。对于FPIT-block，CR-Loss，融合策略等因素进行消融实验，并进行了敏感性分析。

提升方向：1.在编码完成之后进行单步融合，可能限制模型效果；2.加噪过程是否改变图像纹理

[7]Multi-Modal Mutual Attention and Iterative Interaction for Referring Image Segmentation

这篇文章应用于语言指导的图像分割问题，并提出了迭代方式充分利用语言和视觉模态的信息。

在已有的方法中，往往是图像或文本模态占据主导地位，另一种模态则只被使用1到2次。在该工作中，算法迭代式的使用两种模态的信息并进行融合，以实现两种信息的充分交互。



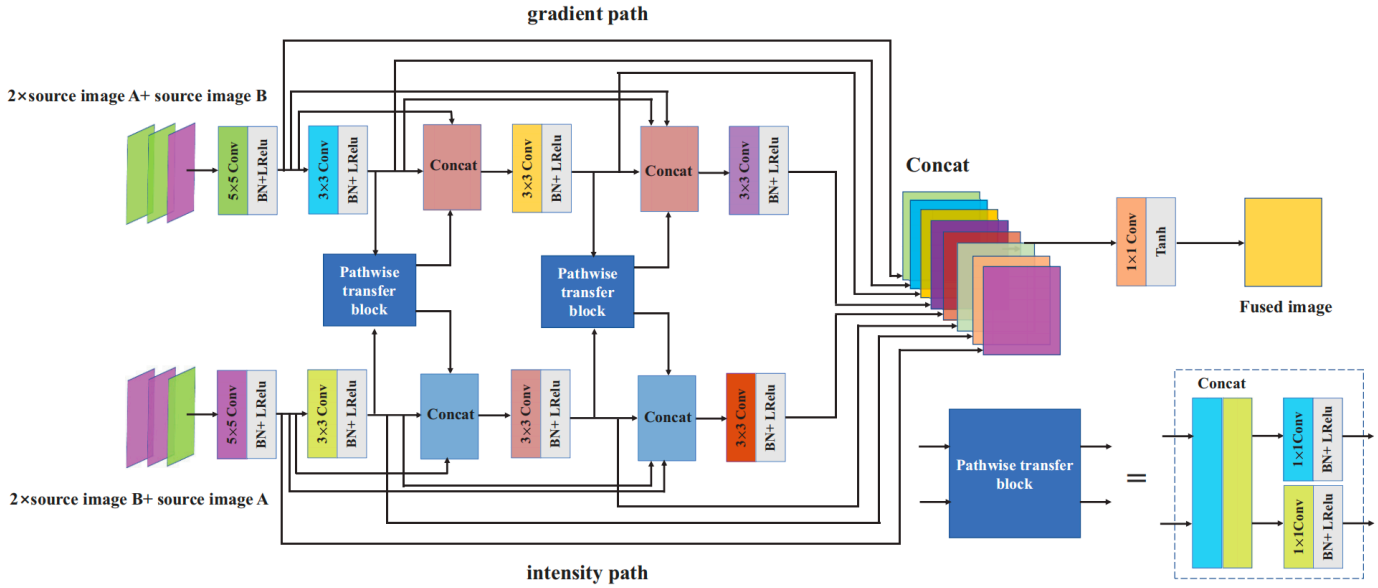
图像信息在经过tranformer编码后，被用于包括Mask Decoder在内的各个decoder输入。与此同时，文本信息也在每层间得到保留。

在实验部分，作者针对算法的每个部分进行消融实验，并与其它sota算法对比。

提升方向：根据作者描述，在面临高度细节化和模糊的表述时会表现不佳。

[8]Rethinking the Image Fusion: A Fast Unified Image Fusion Network based on Proportional Maintenance of Gradient and Intensity

这篇文章在图像融合任务上提出了一个通用的图像融合网络，该网络的出发点为保留梯度和强度信息。此外，文章还设计了一个特殊的损失函数，可以适用于一切图像融合问题。(例如红外和可见光成像，多曝光度图像，医疗图像，多焦段图像等)



该网络结构包含两条信息提取路径，gradient途径包含了texture information，intensity途径包含了intensity information。在每个输入中，都包含在channel上的两种图像的拼接。

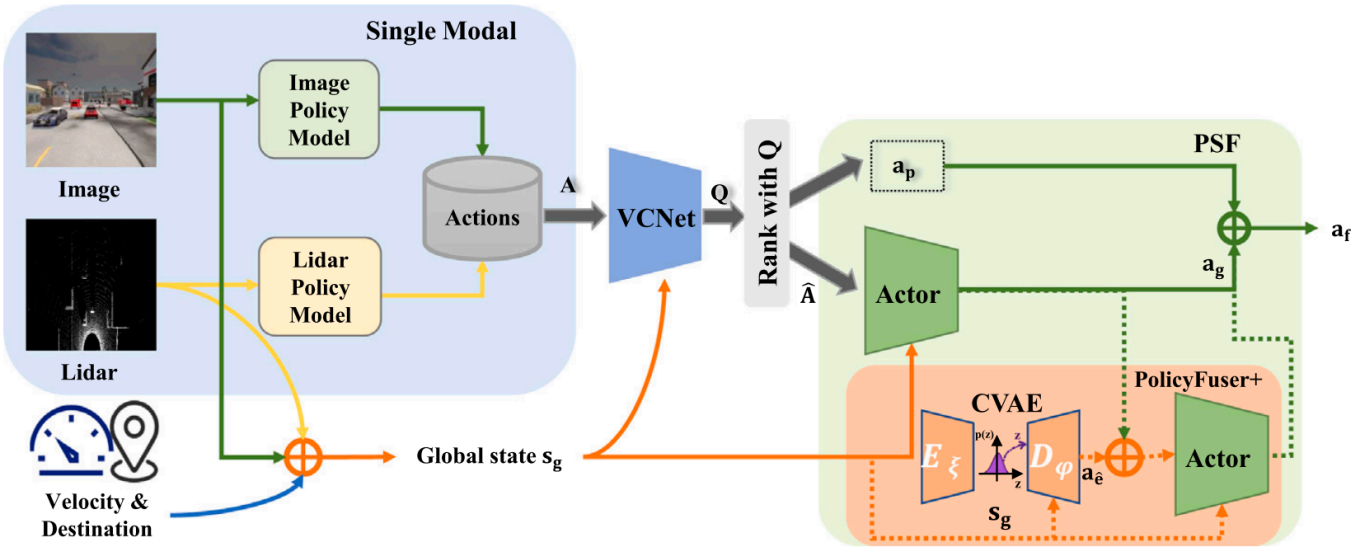
此外，作者还设计了一个通用的损失函数，由gradient loss 和intensity loss组成，可以用于各种图像融合任务。

在实验部分，作者在多种图像融合任务上验证算法可行性。

提升方向：两条path是否应该分别应用loss。

[9]Multi-modal policy fusion for end-to-end autonomous driving

这篇文章的应用场景是利用多模态信息进行的自动驾驶。目前已有的多模态自动驾驶方法在传感器出现故障时会变得极不可靠，而且这些方法在训练过程中也需要复杂的特征对齐或者特征融合神经网络结构。



文章的motivation为：1.降低模型的复杂度；2.提高模型在传感器失灵情况下的可靠性

文章中提出了一个端到端的策略融合自动驾驶算法。该方法利用强化学习选择具有最高Q值的决策作为primary决策，其它的为secondary决策。这些secondary利用primary and secondary policy fusion(PSF)模块fine tune 之前的primary决策。

当所有传感器均运行正常时，模型采用混合决策的方式。而当有一个传感器失灵时，模型则转为采用单传感器的决策模式。

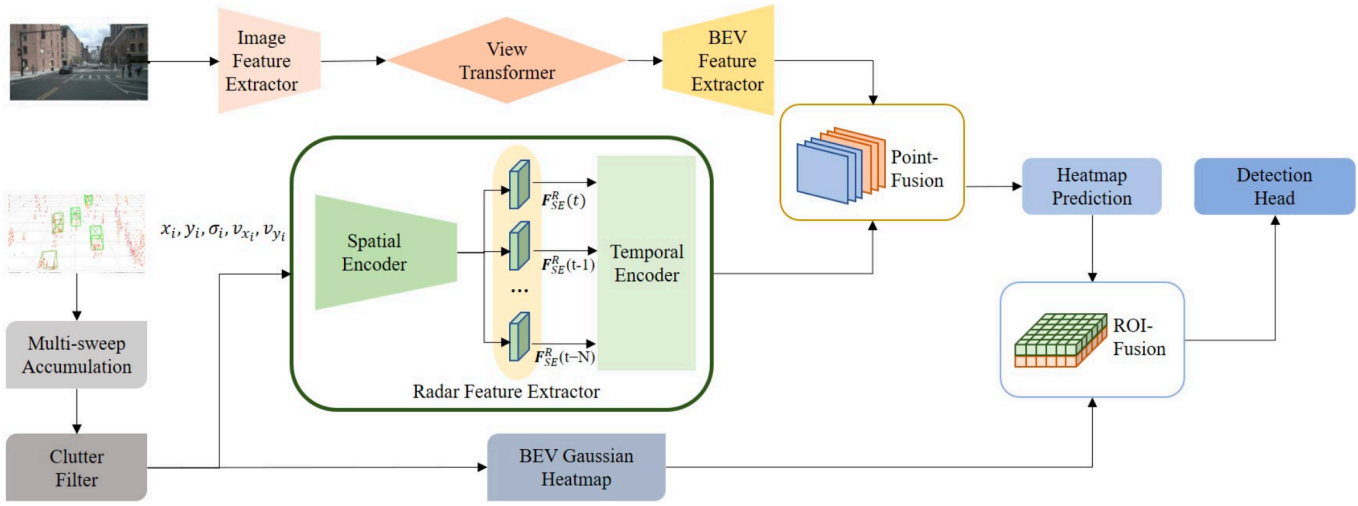
文章的主要贡献为：1.PolicyFuser,结构简单的特征融合端到端的自动驾驶决策算法。2.PSF模块，基于Q值混合primary和secondary决策，使得模型当传感器失灵时依然保持稳定。

在实验部分，作者添加了随机天气作为干扰，并测试了在传感器失灵的情况下的模型表现。

提升方向：1.两个决策网络相对相对简单，不能保证决策的质量。2.inference 和fintune 过程是否能实现实时决策。

[10]Bridging the View Disparity Between Radar and Camera Features for Multi-modal Fusion 3D ObjectDetection

该文章为了解决利用雷达(rader)和摄像数据融合实现3D物体感知的问题，提出了cross-view feature-level fusion framework。该模型先分别提取雷达和图像特征，而后利用两步混合: point-fusion和ROI-fusion获得cross-view图像特征。最终通过anchor-free regression head获得3D感知结果。



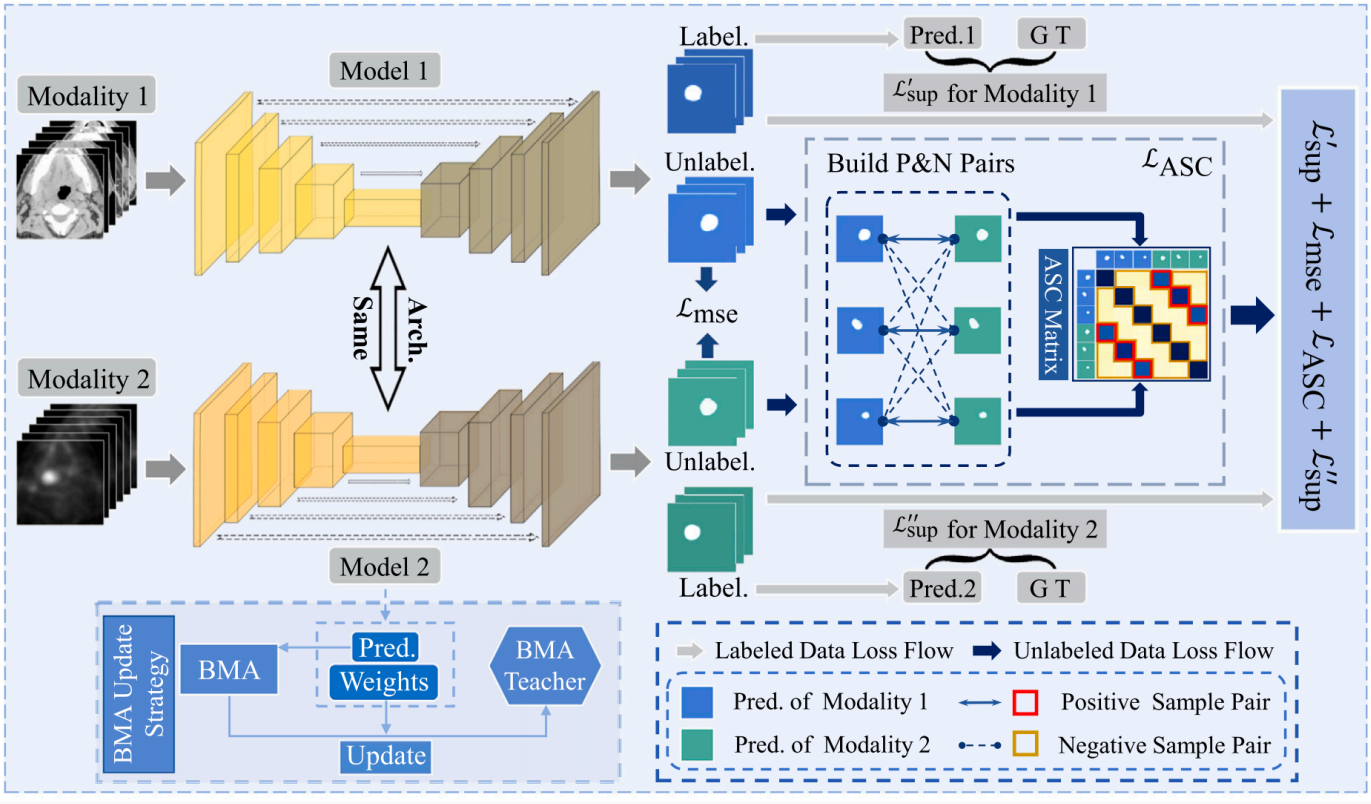
这篇文章实现了高效的特征混合系统，成功克服了雷达和摄影特征之间的视图差异问题。利用雷达时空信息编码器，成功克服了雷达信号稀疏和杂乱的问题，并实现了高效的雷达点特征提取。最终通过两步特征融合方法实现了充分的信息交互，增强和回归的结果。

在实验部分，作者将该方法和其它sota方法进行了对比，并在不同的天气和照明条件下和其它方法进行对比。

提升方向：LSTM结构是否能够及时响应突发情况。

[11]Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation

这篇文章的应用场景是医疗影像的分割问题。这项工作首先解决了在多模态处理医疗影像的问题中，多模态信息需要同时在模型训练和模型推理阶段使用，从而限制了其在临床领域的应用。为了解决这一问题，文章提出了一个半监督的对比学习分割框架semi-supervised contrastive mutual learning(Semi-CML)。该框架在训练过程中可以感知无标注数据之间的关系，并且在推理过程中只使用一种模态。



同时，为了更好的利用多模态信息并提升预测一致性，文章提出了 area-similarity contrastive (ASC) loss，使得一个模态可以学习到来自另一个模态的互补信息。

此外，对于在多模态学习过程中两种模态的表现存在差异的问题，该工作提出了 soft pseudo label re-learning (PReL) scheme来弥补这种差异。

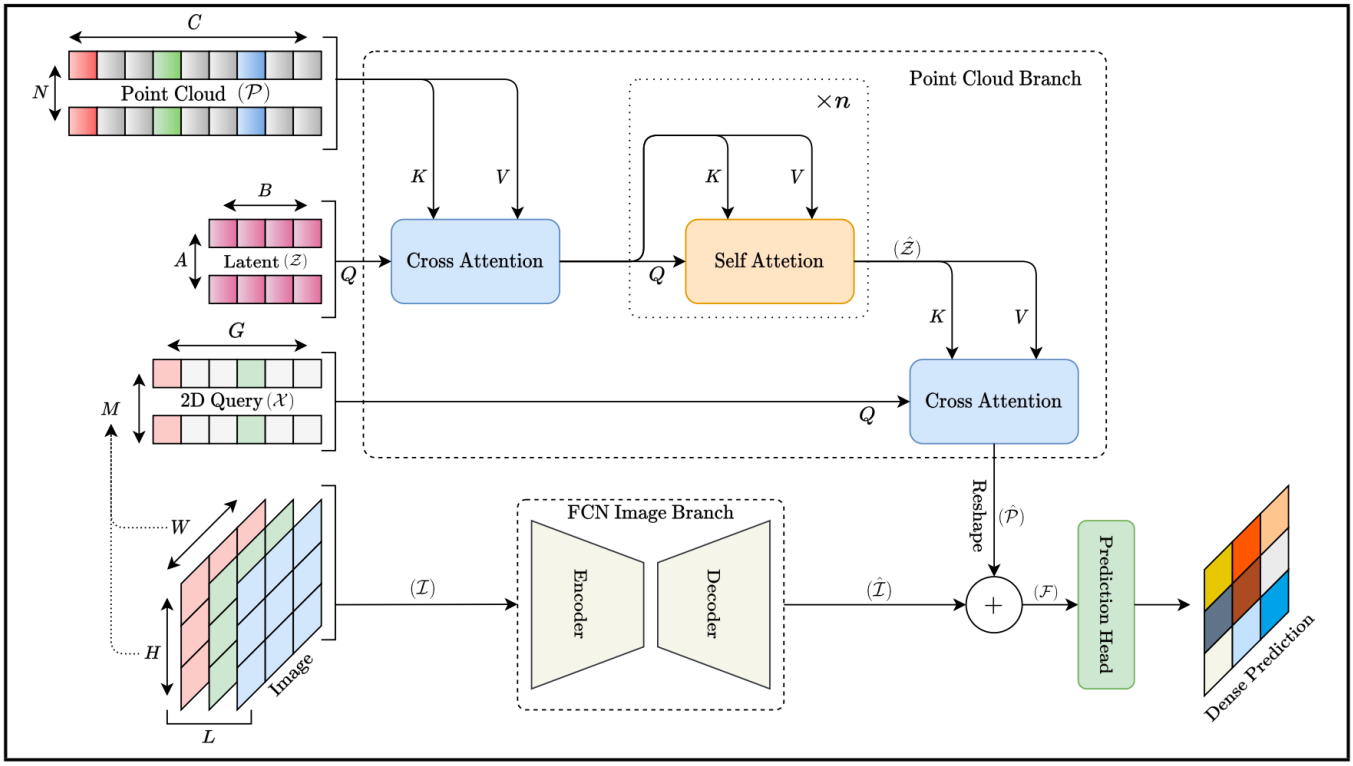
在模型训练阶段，首先输入两个模态成对和非成对的图像，并有ASC loss等进行更新。之后，利用表现更好的模态生成一个best-model moving average (BMA) teacher模型，并在之后利用这个teacher模型改善表现较差的模态。

在实验部分，作者将两个模态的模型和其它sota方法进行了对比，并进行了消融实验。

提升方向：1.ASC loss之外，考虑其他的交互两个模态信息的方式；2.只使用一个模态训练BMA teacher是否不足；3.最终推理进行多模态融合

[12]Transfusion:Multi-modal Fusion Network for Semantic Segmentation

这篇文章尝试解决在多模态输入时（2D彩图和3D点云），不同模态特征间的align困难和可能存在的模态偏差问题。对于语意分割问题，文章直接将图像和点云进行fuse，避免了对于点云的预处理过程中造成的信息丢失。



该方法可以忽略空间稀疏性和变量点密度，对每个样本的点数量没有理论限制，作为端到端可学习的方法可以实现3维到2维特征空间的无缝映射。

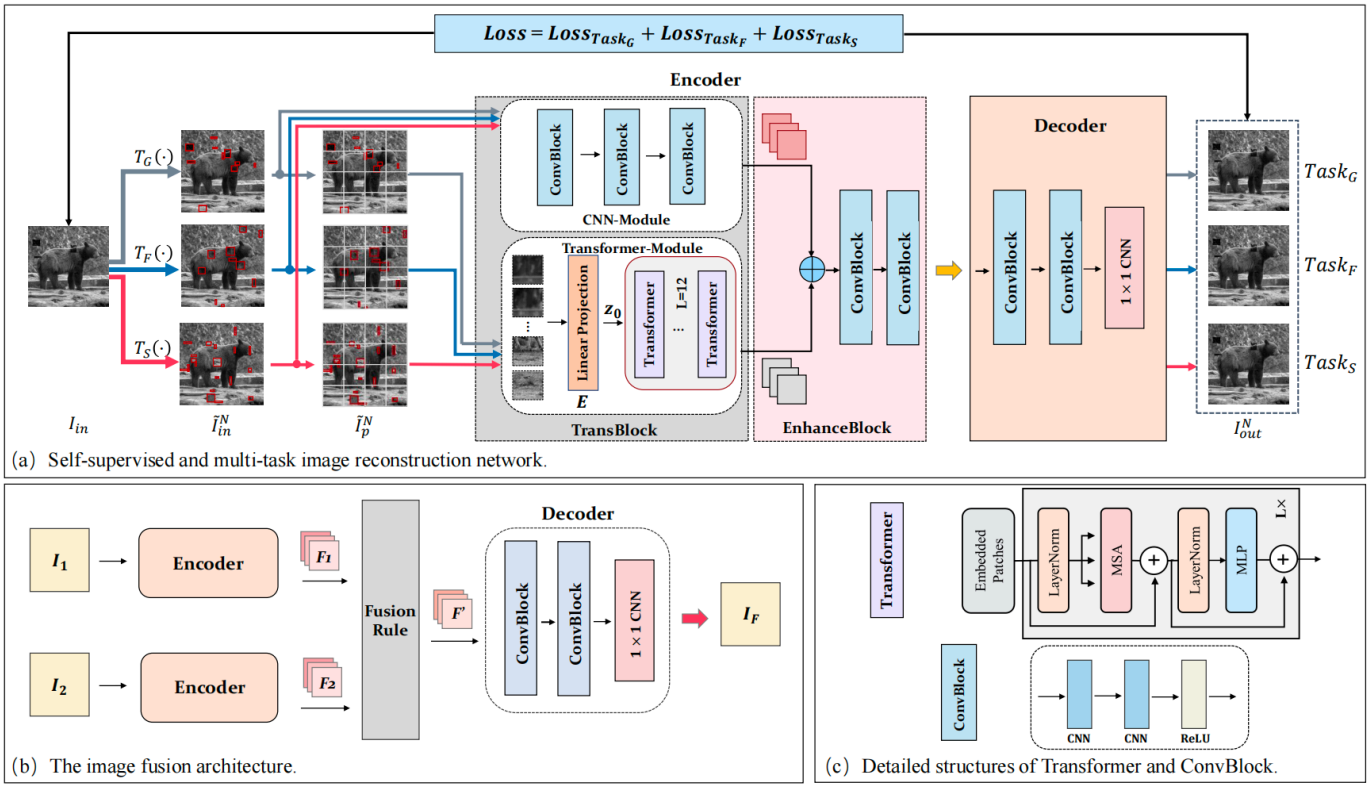
该方法的核心是使用cross attention机制，先编码点云，后编码2维像素坐标，最终利用softmax函数和图像进行混合。

在实验部分，作者和其他baseline进行对比，同时考察了点云特征分支的效果，证明了模态融合的效果。

提升方向：1.提取出来的两种特征融合机制相对简单；2.展示的效果图都较为有序，复杂情况是否可行。

[13]TransMEF:A Transformer-Based Multi-Exposure Image Fusion Framework using Self-Supervised Multi-Task Learning

这篇文章提出了一个基于transformer的多曝光度图像混合框架，利用自监督多任务学习。该网络采用编码-解码结构，不需要ground truth 混合图像。作者设计了三种自监督重建任务，使得模型可以学习多曝光图像的特征并提取普适特征。为了提高网络对于局域和全局信息的感知，编码器部分采取了CNN结合transformer的结构。



利用 **Gamma-based Transformation**，**Fourier-based Transformation**和**Global Region Shuffling** 生成三种任务。利用CNN block 和Transformer block分别提取信息并合成。最后解码得到输出，根据输出生成loss。

在fusion 部分，采用RGB 到YCbCr，在Y channel进行融合的方式。

网络的训练基于一个大型的自然图像数据集。

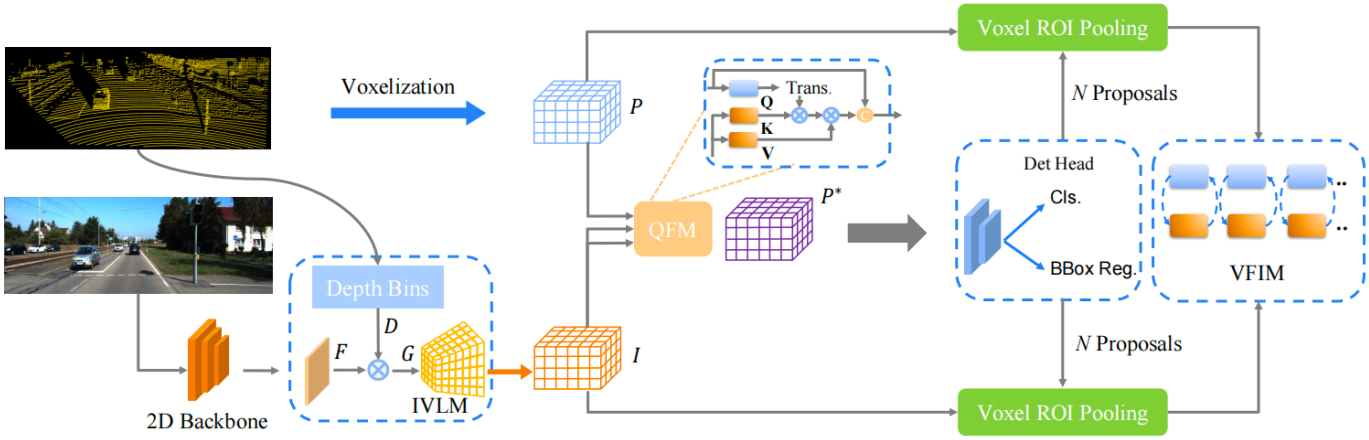
在实验部分，作者将该方法和其他sota方法在主观指标和客观指标上进行对比，并且针对三种任务进行消融实验。

提升方向：将该方法应用于其他类型的任务

[14]Homogeneous Multi-modal Feature Fusion and Interaction

这篇文章提出了一种同时利用图片和点云信息的3D感应算法。

文章利用image voxel lifter module(IVLM)将2维图像特征提升至3维空间，并结合3维体素特征，生成同一化的图像体素特征。之后将图像和点云特征通过query fusion mechanism(QFM)结合。为了增强两个模态在同一物体上的语义一致性，作者引入了voxel feature interaction module (VFIM)。



在VFIM中，算法先在3D detection proposal中取样，然后利用voxel RoI pooling分别提取point voxel 和image voxel的RoI 特征。提取出来的特征经过encoder和mlp后，计算cosine similarity并用于优化模型。

在实验部分，作者在两个数据集上进行实验并对**Query Fusion**，**Multi-modal Feature Structure**和**Voxel Feature Interaction** 进行了消融实验。

提升方向：1.尝试cross attention；

[15]External multi-modal imaging sensor calibration for sensor fusion: A review

这篇文章是关于传感器校准的一篇综述，介绍了各种已有的传感器校准方法。

其中，作者详细介绍了相机-相机校准方法，相机-激光雷达校准方法，相机-雷达校准方法以及其他传感器组合的校准方法。

目前主要的研究方向

不同曝光度图像间的融合

通用的图像融合框架

自然语言引导的图像分割

多传感器信息融合（自动驾驶）

医疗图像融合

克服calibration的困难，对不同传感器的信息不同步/传感器损坏robust

提高多模态情况下的inference效率(降低inference过程中对多模态信息的依赖)

降低对Label数据集的依赖(自监督)

两种模态的充分融合

文献列表

[1]Artifacts Mapping: Multi-Modal Semantic Mapping for Object Detection and 3D Localization

[2]CEKD:Cross-Modal Edge-Privileged Knowledge Distillation for Semantic Scene Understanding Using Only Thermal Images

[3]DBCNet:Dynamic Bilateral Cross-Fusion Network for RGB-T Urban Scene Understanding in Intelligent Vehicles

[4]DeepFusion:Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection

[5]Multi-exposure image fusion via deep perceptual enhancement

[6]Rethinking multi-exposure image fusion with extreme and diverse exposure levels: A robust framework based on Fourier transform and contrastive learning

[7]Multi-Modal Mutual Attention and Iterative Interaction for Referring Image Segmentation

[8]Rethinking the Image Fusion: A Fast Unified Image Fusion Network based on Proportional Maintenance of Gradient and Intensity

[9]Multi-modal policy fusion for end-to-end autonomous driving

[10]Bridging the View Disparity Between Radar and Camera Features for Multi-modal Fusion 3D ObjectDetection

[11]Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation

[12]Transfusion:Multi-modal Fusion Network for Semantic Segmentation

[13]TransMEF:A Transformer-Based Multi-Exposure Image Fusion Framework using Self-Supervised Multi-Task Learning

[14]Homogeneous Multi-modal Feature Fusion and Interaction

[15]External multi-modal imaging sensor calibration for sensor fusion: A review