# MonoFlex论文解读

**分享一篇cvpr 2021上的MonoFlex，MonoFlex网络作为一个延续centernet框架的单目3d检测器，论文的着眼点是优化图片边沿被截断物体的3D检测，同时优化了中心点的深度估计。这个方法也是目前没有extra traning data的条件下，单目3d目标检测在 kitti-car-moderate上结果很好的方法，每帧35ms也比较快了，技巧比较多，非常值得一看的！**

# 主要创新点

主要贡献主要归纳为以下两点：

1. 发现针对截断类(outside object)的目标，从2D到3D映射过程，需要的偏移量的分布与非截断类目标(inside objects)的偏移量分布差别很大，因此要解耦这两类目标，分别进行学习，也就是关注到单目三维目标检测中考虑目标间差异的重要性，提出了截断目标预测的解耦方法。

1. 提出了一种新的目标深度估计公式，它利用不确定性灵活地组合独立的估计器估计对象深度，而不是对所有对象采用单一方法。

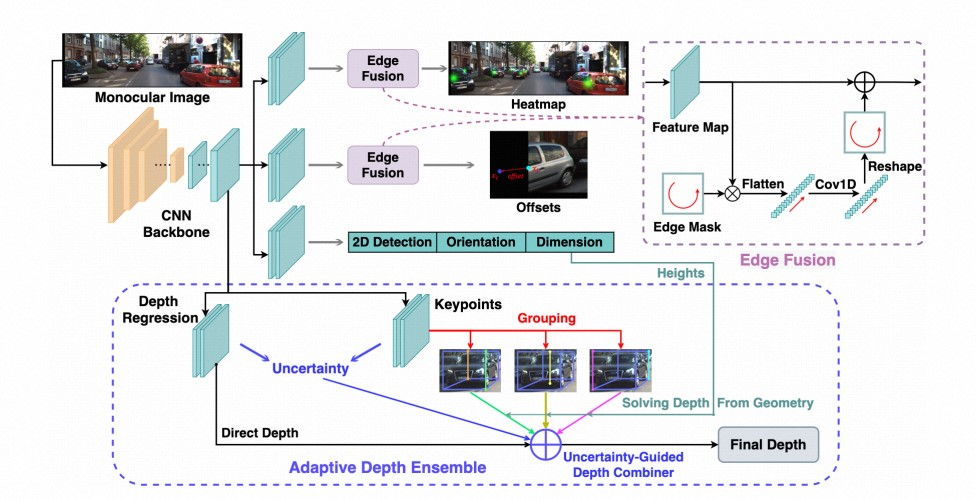
## 主要观点、思考：

目前算法大多数对不考虑目标的截断属性，对正常目标和截断木目标一视同仁，而且对于有遮挡/截断的物体和正常的物体要分开对待。

**考虑到了目标的截断属性，提出了一种灵活的单目3D目标检测框架，使用边缘特征解耦了截断目标和正常目标，分开处理。**在比较重要的深度估计方面，使用了直接预测法、基于关键点的几何计算法，同时结合不确定性，使深度估计的更准确。

# 整体框架

Nonoflex框架以及检测思想是从CenterNet扩展而来的，CenterNet的核心思想是将目标作为一个点，即目标BBox的中心点，检测器采用关键点估计来找到中心点，并回归到其他目标属性，例如2D边界框、维度、方向、关键点和深度。最终深度估计是回归深度和根据估计的关键点和尺寸计算的深度的不确定性组合：



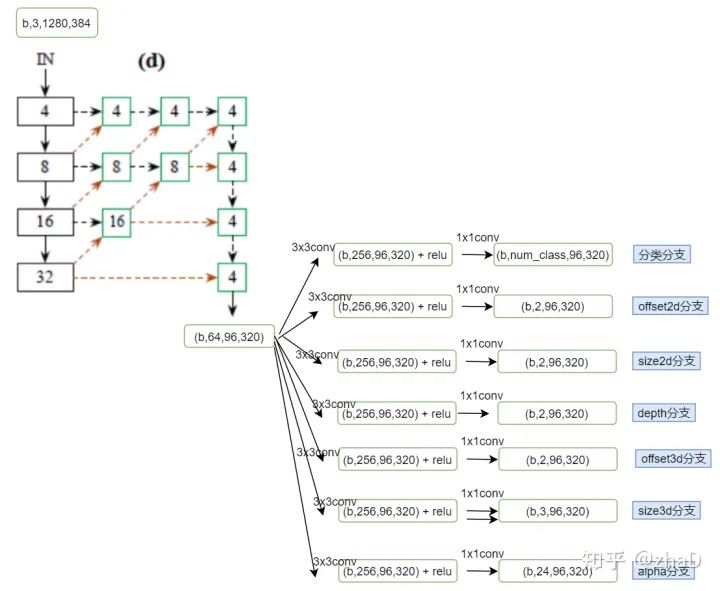
* 首先，CNN主干网络从单张图像中提取特征图作为多个预测头的输入，其中图像级定位涉及热图（Heatmap）和偏移量（Offsets）

* 之后边缘融合（Edge Fusion）模块用于解耦截断对象的特征学习和预测

* 同时自适应深度集合采用四种方法进行深度估计，并同时预测其不确定性，从而形成不确定性加权预测

## 网络总述：

基于centernet的架构，与[smoke](https://zhuanlan.zhihu.com/p/452676265)在head上的差异不大，这里加入了2d模块。



* 分类分支：3dbox中心点投影下来的x\_c、y\_c，为heatmap

* offset2d分支：2dbox中心点与上面3dbox中心点投影下来点的偏移量

* size2d：2dbox的宽高

* depth：相机坐标系下的目标距离z（深度depth），这里加入的不确定度项。

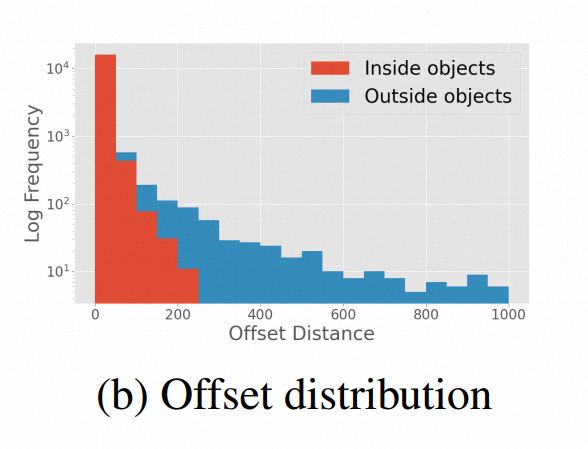
* offset3d：3dbox中心点投影下来的x\_c、y\_c在下采样的过程中的量化误差。可以参考centernet中

* size3d：目标的长宽高。代码中设置的从0开始，而不是从数据集的平均值。

* alpha分支：划分为12个格子来预测。

## 对象解耦:

现有的单目3D检测方法对每个对象使用统一表示，即2D边界框中心点。计算偏移回归以导出投影的3D中心。根据物体的投影3D中心在图像内部还是外部，我们将物体分为两组，内部对象（Inside Objects）和外部对象（Outside Objects）在从2D中心到投影3D中心过程中，呈现完全不同的偏移分布：



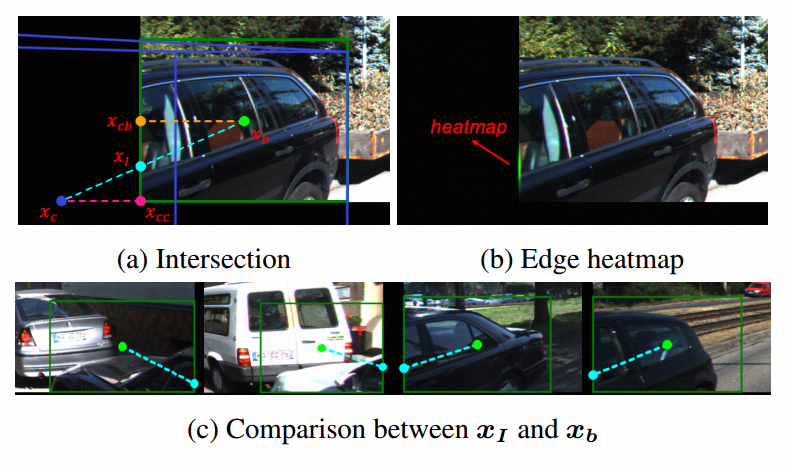
因此，将内外对象的表示和偏移学习进行解耦，**对于3D框中心点投影的预测: 完全在画面里面的obj，直接预测3D框中心点投影及其修正值；对于部分在画面外的obj， 预测2D框中心和（画面外）3D框中心投影的联系与图像边沿的交点及其修正值。**

* 对于投影的3D中心位于图像内部的对象，它们由直接识别，此时的偏移误差如下，其中为CNN下采样率：



* 为了解耦外部对象的表示，通过图像边缘 和 从到的之间的交点来识别外部对象，交点比简单地将或夹持到边界更加有物理意义：



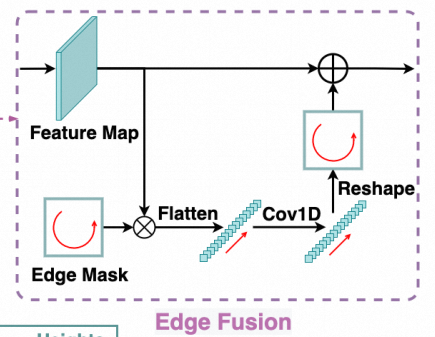


## 边缘融合：

**单独把边沿的像素拿出来做attention，提高offset和heat map的精度。**

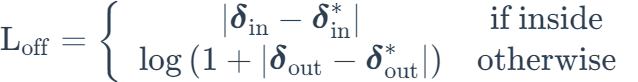
* 尽管内部和外部对象的表示在输出特征的内部和边缘区域中解耦，但共享卷积核仍然难以处理空间变量预测。因此，作者提出了一个边缘融合模块来进一步解耦外部对象的特征学习和预测。

* 如下图所示，模块首先提取特征图的四个边界，并将它们按顺时针顺序（图文不一致）连接成边缘特征向量，然后由两个1×1卷积层处理，以学习截断对象的独特特征。最后，将处理后的向量重新映射到四个边界并添加到输入特征图。当应用于热图预测时，边缘特征可以专门预测外部对象的边缘热图，从而不会混淆内部对象的定位。为了回归偏移，和之间的显著尺度差异可以通过边缘融合模块解决。



## Loss：

作者采用L1 Loss回归，Log-Scale L1 Loss回归，因为它对极端异常值更加鲁棒，偏移损失计算为：



其中，和表示预测，和表示GT。

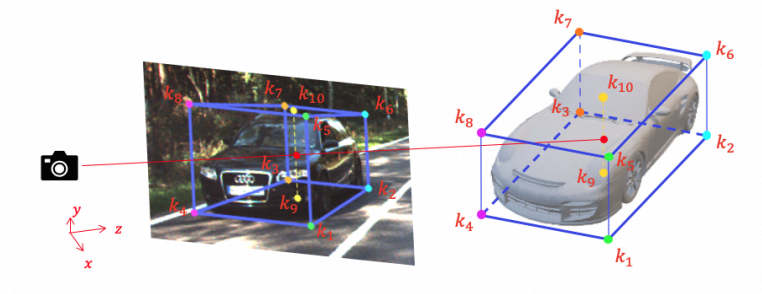
## 视觉特征的回归:

视觉属性的回归，包括对象的2D边界框、尺寸、方向和关键点。

### 深度估计

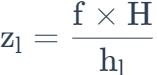
#### 关键点估计：

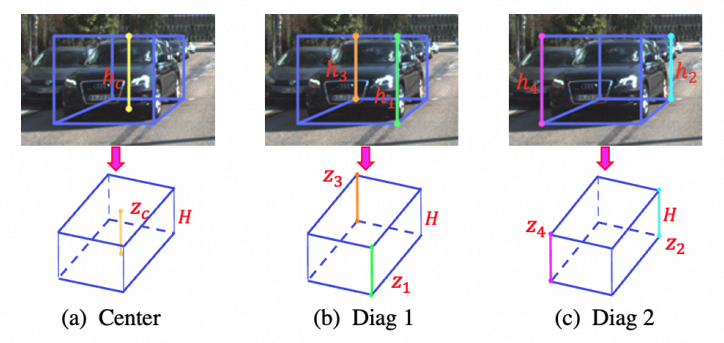
为每个对象定义10个关键点，其中包括3D边界框的8个顶点，i= 1...8、底部中心和顶部中心的投影：



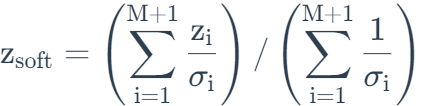
在预测几何深度的时候，会利用到这十个顶点，来计算三类目标的2D高度，再根据投影公式来计算目标的深度。

#### 自适应深度集成：

利用目标的2D高度和3D高度之间的相对比例来计算目标深度（即投影公式：），从高度求解深度不仅与orientation估计无关，而且受dimension估计误差的影响较小（预测时直接回归10个关键点，而计算targets中的10个关键点时则是需要利用labels中的目标三维坐标、偏航角以及长高宽）：



利用M+1个深度估计值和不确定性，进行加权求解平均值，集成方式如下：



soft ensemble可以为那些更自信的估计器分配更多的权重，同时对潜在的不准确不确定性拥有一定鲁棒性。

# 主要结构组成:

1. 特征提取模块:这个部分负责从输入图像中提取有用的特征表示。通常采用预训练的卷积神经网络(CNN)作为特征提取器，例如VGG、ResNet或EfficientNet等。这些网络通常包含多个卷积层和池化层，能够有效地捕捉图像中的语义和局部特征。 2. 物体检测头:物体检测头负责在图像中检测和定位物体。它通常由卷积层和全连接层组成用于生成物体边界框的位置和类别预测。常用的物体检测算法，如YOLO、SSD或Faster R- CNN等，可以作为MonoFlex的物体检测头。 3. 2D边界框重定位模块:这个模块用于提高物体位置的准确性。它使用几何推理和约束来对检测到的物体的2D边界框进行重定位。通过分析物体在图像中的形状、大小和位置等信息，结合先验知识和模型，能够更准确地估计物体的位置。 4. 深度估计模块:这个模块用于推断物体的深度信息。它可以基于视差(disparity)或其他深度估计算法，通过学习图像的视差信息或结合其他传感器(如深度相机)提供的深度数据，估计物体的深度信息。 5. 姿态估计模块:这个模块用于估计物体的姿态，包括旋转角度和方向等。通过利用物体的边界框和深度信息，结合几何变换和计算机视觉技术，可以推断物体的姿态。 整个MonoFlex网络结构中的这些部分通常会联合进行训练，通过端到端学习来优化整个系统。这样可以使不同部分之间的信息流动更加顺畅，并提高单目三维物体检测的准确性和鲁棒性。