工作总结与计划

一、 工作总结（按时间划分）

日期	工作内容			完成情况及主要问题
	上午	下午	晚上	
星期一	读论文	准备汇报	开组会	对论文中不清楚的的一些点有了新的理解，如深度估计中年的尺度问题
星期二	修改代码	修改代码和训练	审阅论文	按照学长说的问题改进后训练
星期三	审阅论文和训练	读论文	读论文	阅读深度估计和语义分割联合学习的论文
星期四	读论文	读论文	休息	完成论文的阅读
星期五	写代码	处理杂事	写代码	写 v1 和 v2 的代码
星期六	写周报	写周报	写代码	完成周报

二、下一步计划（按任务划分）

编号	工作内容	目标	相关配合
1	训练 v1	2 月 3 日前完成	无
2	确认 v2 的可行性，完成编写代码	2 月 7 日前完成	无
3	想 v3 两个任务在 decoder 的交互方式	2 月 7 日前完成	无
4			
5			
主要风险	无		

三、个人分析与总结

内容提要	
1	进度方面：在代码完成方面效率较低，把重点放在完成自己的模型和代码上

2	课题方面：本周阅读的论文在两个任务的交互上比起仅通过损失函数约束有所增强
3	
4	

本周了解了在深度图像超分辨率重建任务中输入数据的方法，包括对原图像的裁剪成等大的 patch，并尽可能包括图像的所有信息而不要随机裁剪，此外在输入模型之前可以记录 patch 深度值的最大值，并将深度图像归一到[0,1]区间内，并在模型输出后结合记录的最大值计算损失，可以帮助模型更好的训练。

模型方面，本周先简单地将 PMBANet 和 CLIFFNet 的模型通过损失函数结合起来，但两个任务的结合方式并不紧密，因为 PMBANet 的彩色引导分支没有移除，周日拿到卡之后开始训练一下。另外根据之前读的论文设计了 v2 的模型，v2 在首先分别对高分彩色图像和插值后的深度图像提取特征，然后对特征进行融合，并分别送入深度估计解码器和超分解码器，两个解码器都是渐进提高分辨率的，在不同分辨率对特征进行损失函数约束。在 v3 希望将这种约束方式从损失函数改进为其他模块。

本周阅读的论文为联合深度估计与语义分割两个任务的论文，相较于之前联合学习的论文，本文不是通过损失函数来对任务进行约束，而是通过分析深度估计与语义分割内在的关联性，引入语义对象（semantic objectness）的概念从对立任务中挖掘潜在特征，从而促进两个任务的性能提升。

四、论文总结

论文标题	SOSD-Net: Joint Semantic Object Segmentation and Depth Estimation from Monocular images
作者及单位	Lei He[a,b], Jiwen Lu[b], Guanghui Wang[c], Shiyu Song[a], Jie Zhoub,[d,*] [a]Baidu Autonomous Driving Technology Department (ADT) [b]Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing, China [c]Department of Computer Science, Ryerson University, Toronto, ON, Canada M5B 2K3 [d]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
论文出处	2021-arXiv
创新点提炼	1. We propose a Semantic Objectness Segmentation and Depth Estimation Network (SOSD-Net) to enhance the learning ability of joint monocular depth estimation and semantic segmentation. 2. An effective learning strategy is proposed to alternatively update the specific weights of SOSD-Net, which significantly improves the performance of the two tasks. 3. We achieve competing results over the state-of-the-art one-stage models on two popular benchmarks.
个人想法	本文通过挖掘深度估计与语义分割的内在关联性，分别通过深度到语义单元和语义到深度单元将深度/语义特征的隐含特征结合到语义分割/深度估计任务中，从而促进两个任务的性能提升。可以在 v3 的模型中，设计类似的模块，在解码过程中将深度估计的特征用于深度图像的超分辨率重建，而不是仅仅通过不同分辨率的约束。

论文方法及结论：

1. 论文提出的问题

深度估计和语义分割在场景理解中起着重要作用。目前采用多任务学习的方法侧重于共享公共特征或从

相应的分支拼接特征图。然而，这些方法缺乏对几何线索和场景解析的相关性的深入考虑。

2. 解决的方法

本文首先通过对成像过程的分析，引入语义对象（semantic objectness）的概念来挖掘这两个任务之间的几何关系，然后提出了基于对象假设的语义对象分割和深度估计网络(SOSD 网络)。据我们所知，SOSD 网络是第一个利用几何约束同时进行单目深度估计和语义分割的网络。此外，考虑到这两个任务之间的相互隐含关系，我们利用期望最大化算法的迭代思想来更有效地训练所提出的网络。

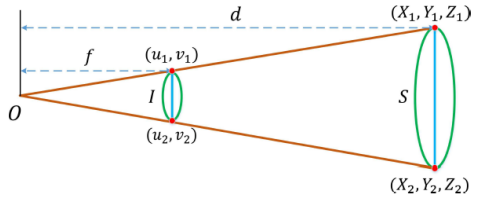


Figure 2: The projection process of a planar object. Where O is the optical center, I is the image of the planar object S, d is the depth of the object, and f is the focal length.

假设空间物体是线性的，如图所示。根据

$$d\begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_x \\ 0 & f_y & u_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X_1 \\ Y_1 \\ d \end{bmatrix}$$

透视投影模型，平面空间物体 S 在(f, O)下的图像为 I

其中(X_1, Y_1)是空间点的坐标，(u, v)是图像上空间点的坐标，(u_x, u_y)是焦点的坐标，(f_x, f_y)对应于相机的焦距。

单目深度估计是一个不适定问题，准确的深度难以恢复。但如果只考虑物体层面的深度，假设物体内部区域的深度近似一致，那么几何关系可以简化为物体的以下 2D-3D 尺寸信息。

$$\Delta u = \frac{f_x \Delta X}{d},\ \Delta v = \frac{f_y \Delta Y}{d}$$

其中，$\Delta u = u_1 - u_2$，$\Delta v = v_1 - v_2$，$\Delta X = X_1 - X_2$，$\Delta Y = Y_1 - Y_2$。可以将上述 2D-3D 尺寸信息扩展到 2D-3D 区域信息。

$$d^2 = \frac{f_x f_y \Delta X \Delta Y}{\Delta u \Delta v}$$

上式的几何关系在本文中被称为语义对象（semantic objectness），其嵌入了语义和相应深度的相关性。在语义分割之后，我们可以通过简单的后处理来获得对象的 2D 区域信息$\Delta u \Delta v$。另外，特定透视下物体的面积信息$\Delta X \Delta Y$是唯一的。因此，我们可以在对象级语义和相应的深度之间建立密切的关系。

实际上，大多数物体内部区域的深度是不一致的。然而，如果仅考虑物体的局部区域，则满足一致深度的假设。目前的深度神经网络由于其非线性和大量的参数可以表达非常复杂的函数。因此，我们利用这一强有力的工具来表达局部隐含关系，并引入 SOSD 网络来嵌入单目深度估计和语义分割之间的语义对象关系。

SOSD 网络由四个部分组成:用于提取上下文特征的 CNN backbone、用于三个特征图(公共表示、语义特征、深度特征)的解码器、用于学习单目深度的语义到深度单元、以及用于学习语义分割的深度到语义单元。

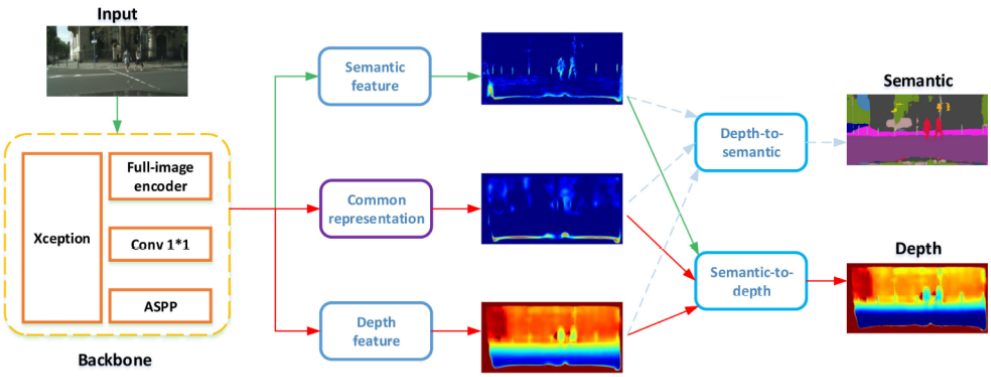


Figure 3: Our proposed SOSD-Net architecture leverages a shared-encoder backbone and a Decoder for semantic feature, common representation and depth feature, followed by depth-to-semantic and semantic-to-depth modules to learn semantic segmentation and depth estimation from a single image, respectively.

Backbone 输入一个彩色图像，并生成一个中间特征图，供每个子任务处理，由 exception-65 和三个并行组件组成，即：ASPP，跨通道学习器（CNN 1*1）和一个全图像提取器。在来自全图像提取器的全局信息的指导下，用 ASPP 纯 1×1 卷积来有效地融合复杂的上下文信息。

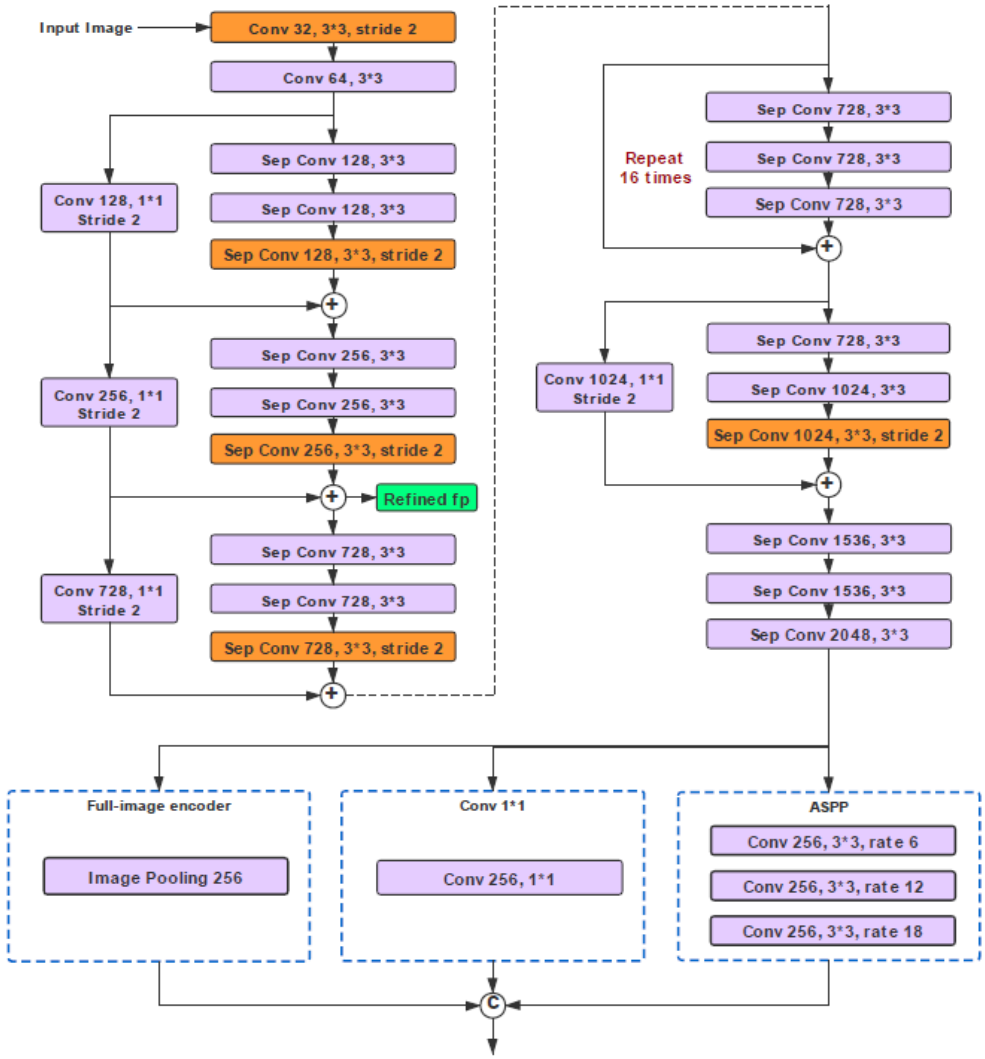
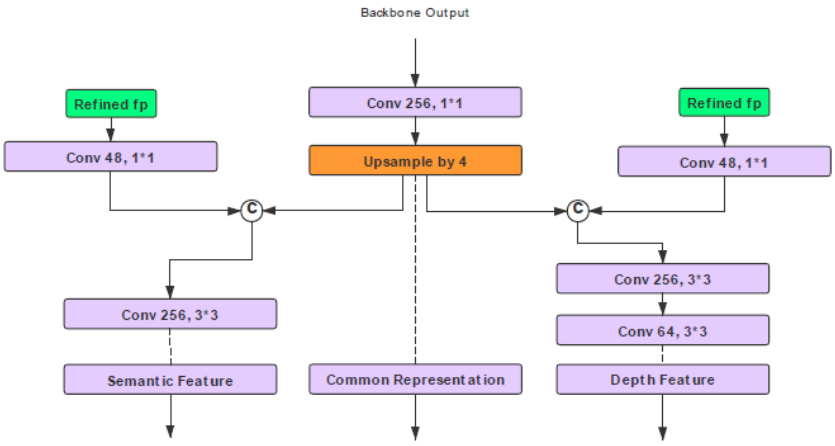


Figure 4: The detailed structure of the Backbone.

基于来自backbone 的全局特征图，解码器的作用主要是分别提取语义特征、公共表示和深度特征。为了弥补步幅卷积造成的结构损失，解码器融合了 backbone 的细化 fp(绿色块)。基于细化后的 fp，首先用一个卷积层分别为每个任务提取信息。然后，结合上采样全局特征映射，解码器使用一个卷积层和两个卷积层来分别生成语义特征和深度特征。



structure of the Decoder. The Refined fp (green block) is generated

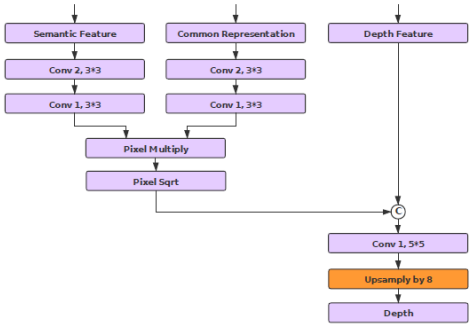


Figure 6: The structure of the semantic-to-depth module.

在获得公共特征之后，用于单目深度估计的经典解码器使用跳跃连接和上采样模块来获得高分辨率深度图。为了嵌入用于单目深度估计的语义对象信息，提出了语义到深度模块来有效地融合深度 2D-3D 区域信息。如图所示，深度三维区域信息是从通用特征图中提取出来的，定义为某个物体在某个视角下的深度区域信息。

在语义到深度单元，首先利用两个具有 2 和 1 通道的卷积层来生成热图，该热图被称为与$\Delta X \Delta Y$相关的对象的 3D 潜在共享表示，然后采用另外两个具有 2 和 1 通道的卷积层来从语义分割中获得另一个热图，该热图是与$\Delta u \Delta v^{-1}$相关的对象的 2D 潜在共享表示。由于公共数据集的焦距是固定的，我们使用批量归一化将焦距信息自动嵌入到两个子分支中。接下来，我们利用深度 2D-3D 区域特征，通过利用像素乘法和平方根运算来推断深度线索。在结合来自深度特征和 Pixel Sqrt 的信息之后，该模块进行具有 1 个通道的卷积层以推断深度图。最后，我们对先前的深度采用上采样操作(双线性插值)来获得全分辨率深度。

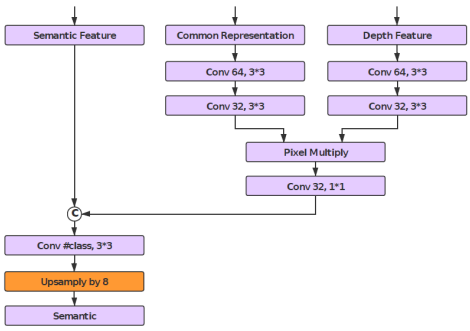


Figure 7: The structure of the depth-to-semantic module.

类似语义到深度，语义分支还通过集成来自先前的特征来嵌入深度 2D-3D 区域特征。首先用两个具有 64 和 32 通道的卷积层，从与d^{-2}相关的深度分支生成一个潜在变量。至于物体的 3D 潜在共享表示，我们通过在公共表示上利用另外两个具有 64 和 32 通道的卷积层来实现。在通过逐像素乘法融合来自两个子分支的信息之后，我们使用 32 个通道的 1×1 卷积来解析语义线索。该模块将语义特征的特征图和语义线索的连接起来，增加了一个卷积层来推断语义分割。为了获得全分辨率分割，我们采用相同的上采样操作来提高语义分割的分辨率。

3. 实验结果结论

Method	Segmentation mIoU [%]	Disparity error [px]
Kendall [18]	64.2	2.65
GradNorm [19]	64.8	2.57
Ozan [20]	66.6	2.54
ESOSD-Net	**68.2**	**2.41**

Table 4: Performance of the multi-task algorithms in semantic segmentation and depth estimation on the CityScapes dataset (sub-sampled to a resolution of 256×512).

Method	samples	**rel**	**rms**	**log$_{10}$**	δ_1	δ_2	δ_3
Two-stage:							
Joint HCRF [15]	795	0.220	0.745	0.094	0.605	0.890	0.970
Jafari *et al.* [65]	795	0.157	0.673	0.068	0.762	0.948	0.988
PAD-Net [54]	795	0.120	0.582	0.055	0.817	0.954	0.987
One-stage:							
Make3D [41]	795	0.349	1.214	-	0.447	0.745	0.897
DepthTransfer [7]	795	0.35	1.20	0.131	-	-	-
Liu *et al.* [66]	795	0.335	1.06	0.127	-	-	-
Li *et al.* [67]	795	0.232	0.821	0.094	-	-	-
Liu *et al.* [68]	795	0.230	0.824	0.095	0.614	0.883	0.975
Wang *et al.* [15]	795	0.220	0.745	0.094	0.605	0.890	0.970
Eigen *et al.*[11]	120k	0.215	0.907	-	0.611	0.887	0.971
R. and T. [69]	795	0.187	0.744	0.078	-	-	-
E. and F. [4]	795	0.158	0.641	-	0.769	0.950	0.988
He *et al.* [12]	48k	0.151	0.572	0.064	0.789	0.948	0.98
Lai [44]	96k	0.129	0.583	0.056	0.811	0.953	0.988
DORN [5]	120k	**0.115**	**0.509**	**0.051**	**0.828**	**0.965**	**0.992**
ESOSD-Net	795	0.145	0.514	0.062	0.805	0.962	**0.992**

Table 6: Quantitative comparison with state-of-the-art methods on the depth estimation task on the NYU Depth v2 dataset (480×640).

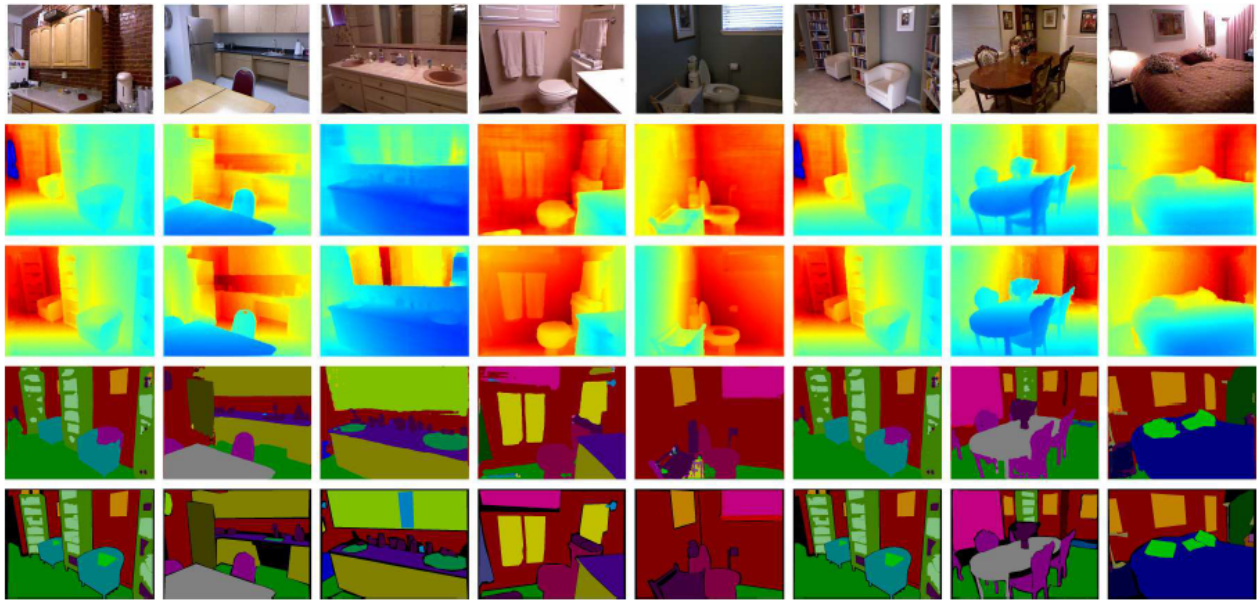
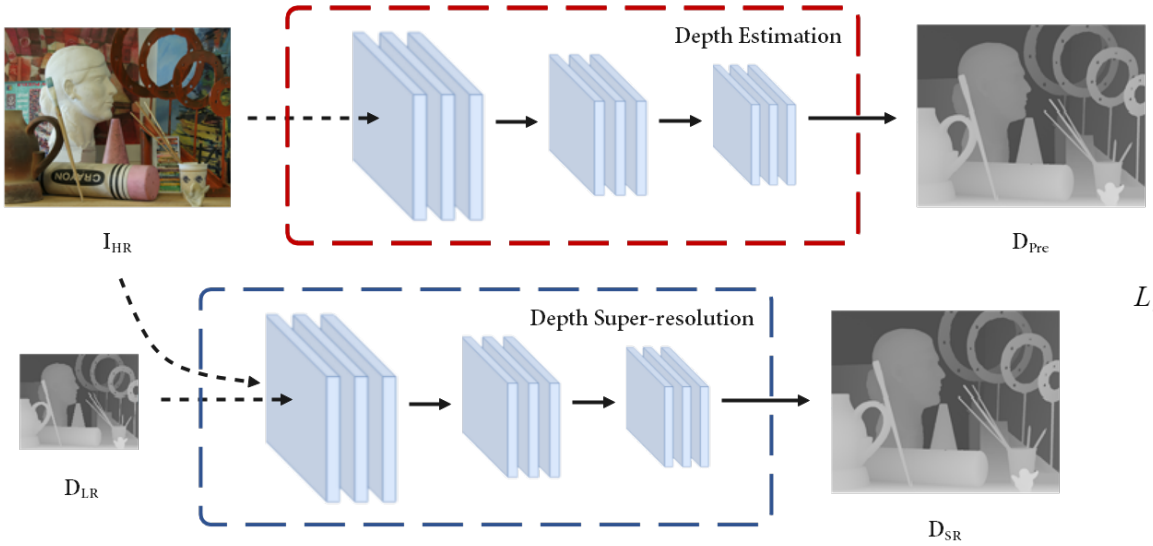


Figure 10: Qualitative examples of monocular depth estimation and 40-classes scene parsing results on the NYU Depth v2 dataset (480×640). The second and the fourth rows corresponding to the predictions of the depth estimation and semantic segmentation. The third and the last rows corresponding to the ground truth of the depth estimation and semantic segmentation, respectively.

五、模型

V1

将 PMBANet 和 CLIFFNet 的模型通过损失函数结合起来，但两个任务的结合方式并不紧密，因为 PMBANet 的彩色引导分支没有移除，仅通过损失函数（包括重建损失、估计损失和深度图一致性损失）将两个任务联合起来，没有利用任务的相关性。



$$L_{total}=\alpha L_{mde}+\beta L_{sr}+\gamma L_{con}$$

V2

首先分别对高分彩色图像和插值后的深度图像提取特征，然后对特征进行融合，并分别送入深度估计解码器和超分解码器，两个解码器都是渐进提高分辨率的，在不同分辨率对特征进行损失函数约束。在 v3 希望将这种约束方式从损失函数改进为其他模块。