汕頭大學
SHANTOU UNIVERSITY

基于注意力机制的室内RGBD图像语义分割

报告人：邱本章

汕头大学人工智能与机器人实验室

指导老师：范衡

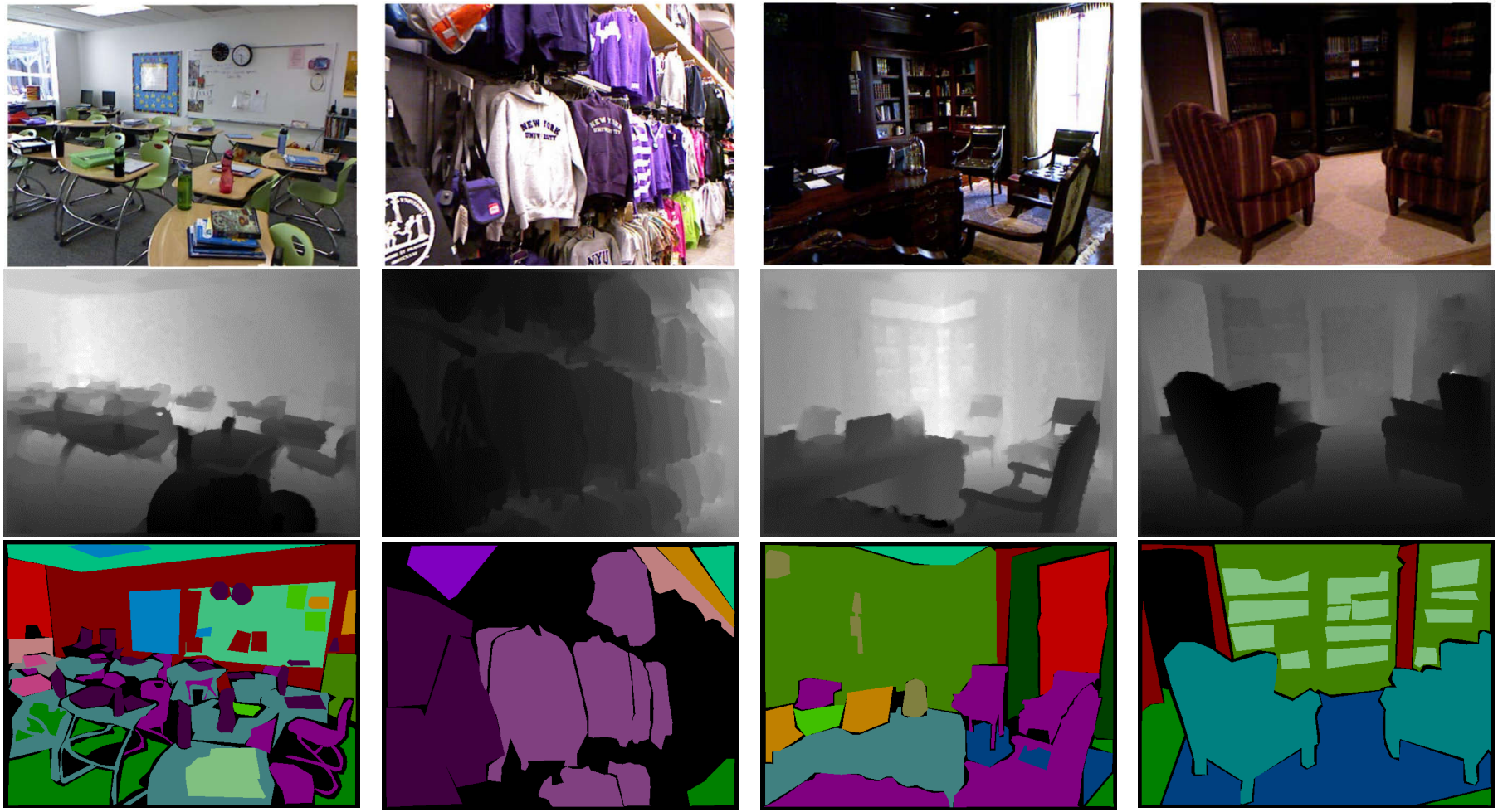
2020/6/12

Contents

研究背景

- 语义分割广泛应用于图像处理中，传统语义分割通常只采用RGB图像；
- 现在机器人和自动驾驶汽车上搭载有RGBD相机（如：Kinect、Real Sense）；
- RGBD数据集（如NYU V2、SUN）的增多；
- 利用RGB图像无法捕捉到的深度信息，结合深度图像可以显著提高分割精度。

研究背景 研究难点



NYU V2 数据集示例

研究背景　研究难点

（1）多视角、多场景、多类别；

（2）光照变化较大，存在高曝光、弱光场景，以及光线变化场景；

（3）互相遮挡，复杂而广泛的上下文关系；

（4）严重的类别不均衡；

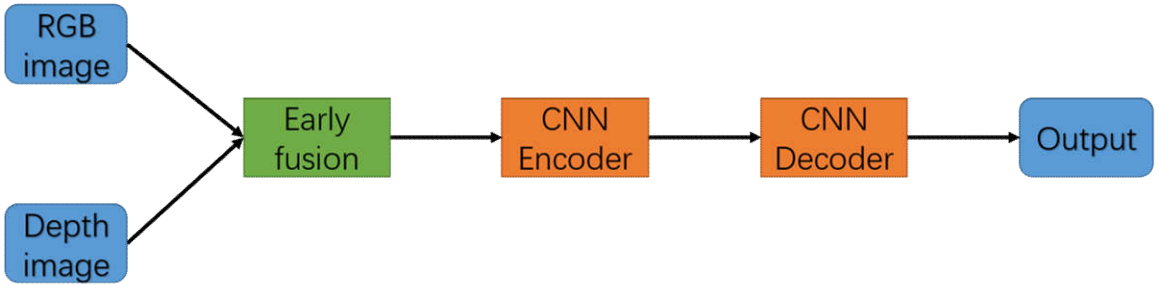
（5）存在很多小物体以及难识别物体（镜子）；

（6）数据存在一些漏标记和错标记情况。

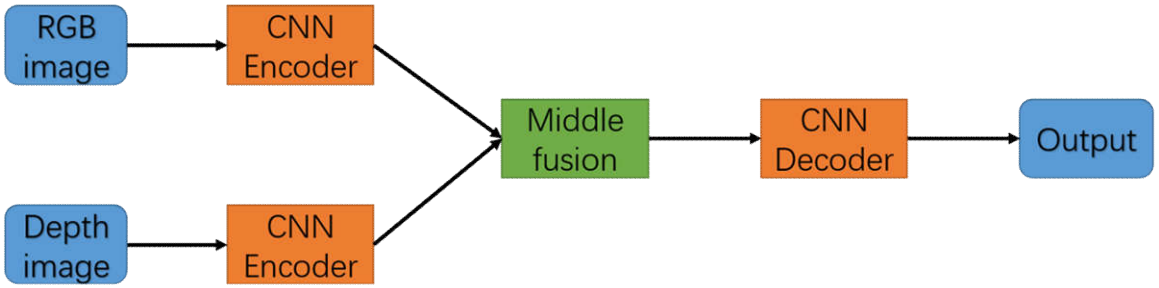
如何将RGB图像与深度图像有效结合，是RGBD图像语义分割研究的重点。

NYU V2 数据集示例

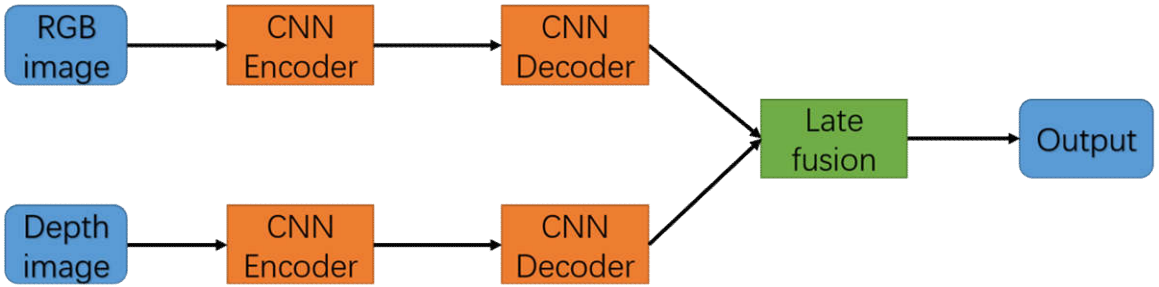
相关工作 RGBD语义分割



(a) 早期融合



(b) 中期融合

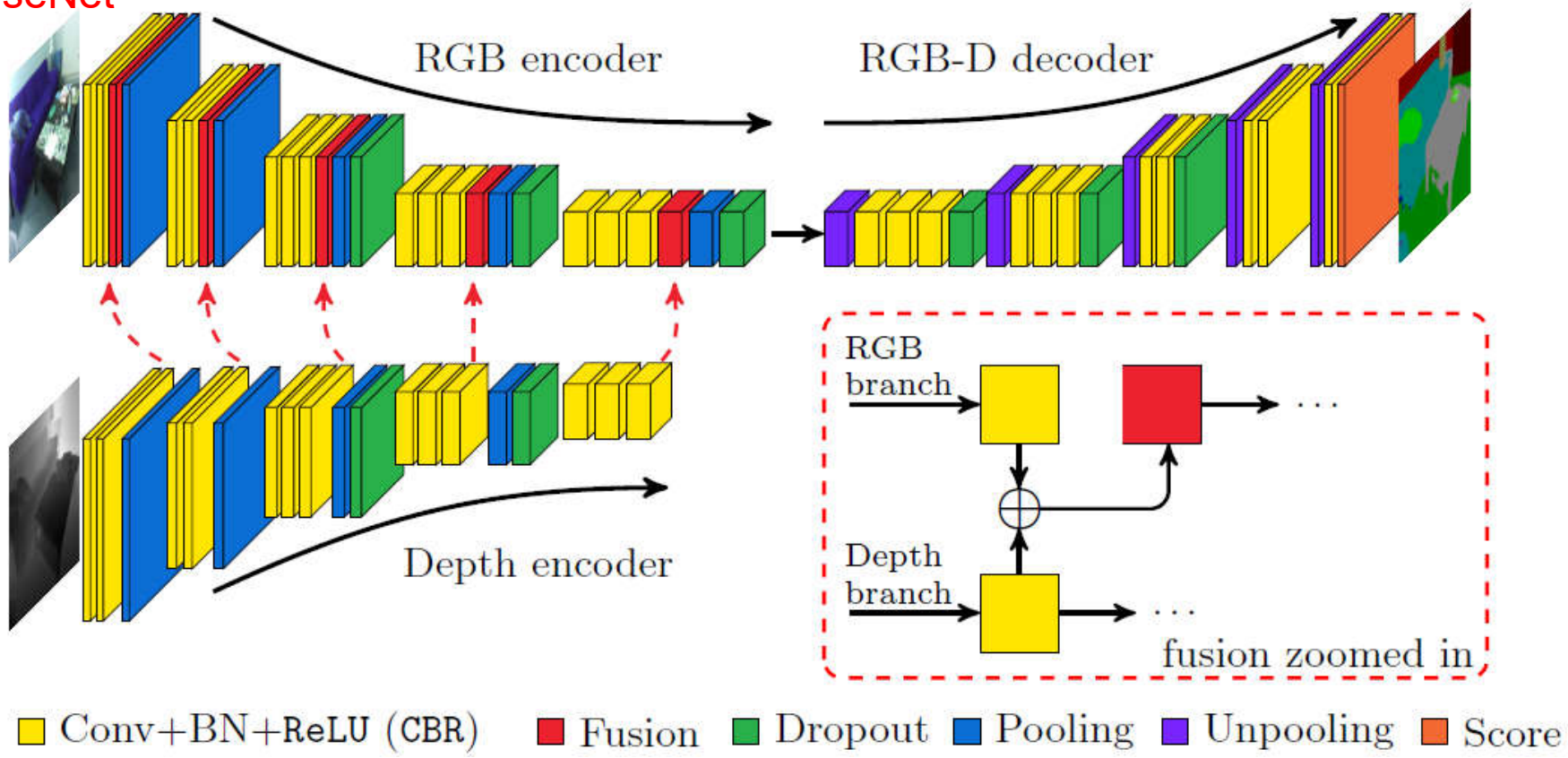


(c) 晚期融合

常见RGB图像与深度图像融合方式

相关工作 RGBD语义分割

FuseNet

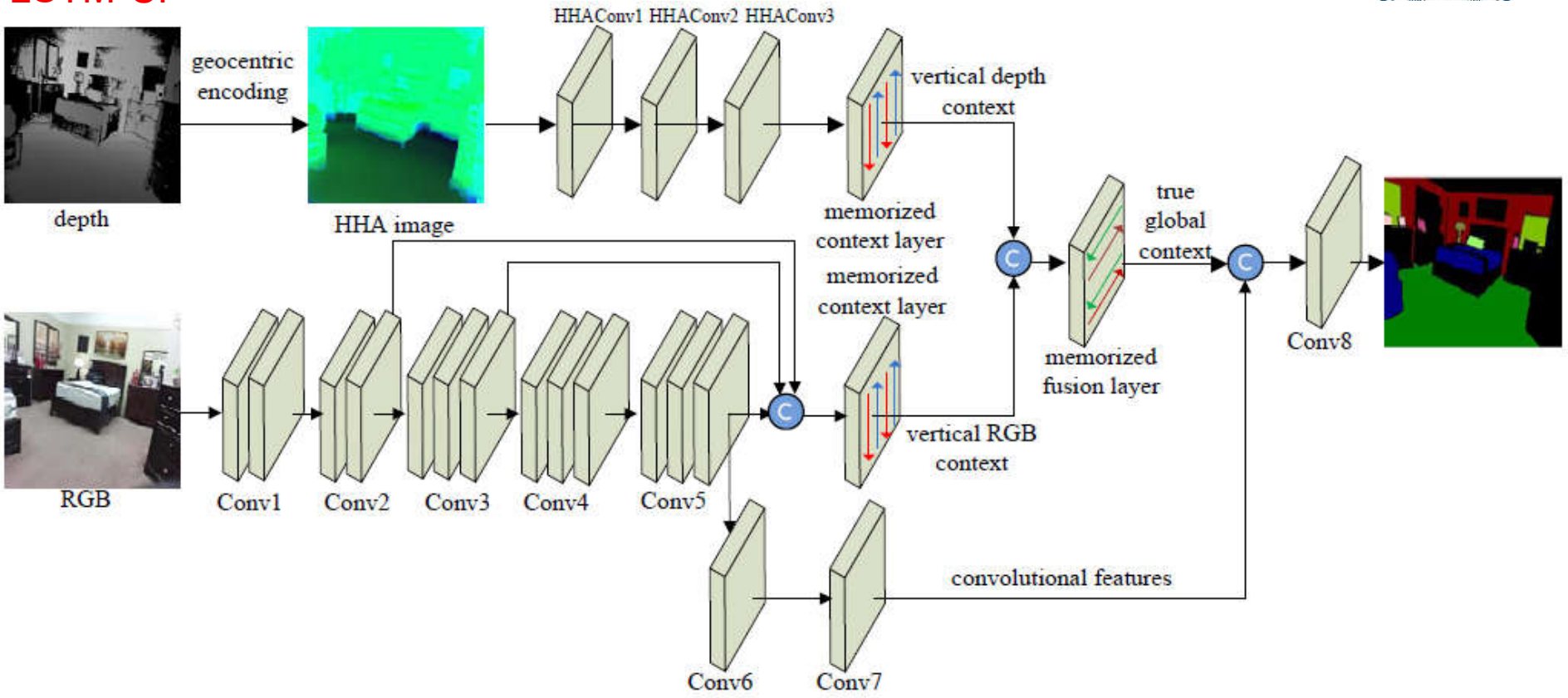


将深度通道中的特征映射逐层叠加进RGB通道中特征映射，实现多层级特征融合，编码器采用VGG-16结构。

[1] Hazirbas, Caner, et al. "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture." *Asian conference on computer vision*. Springer, Cham, 2016.

相关工作 RGBD语义分割

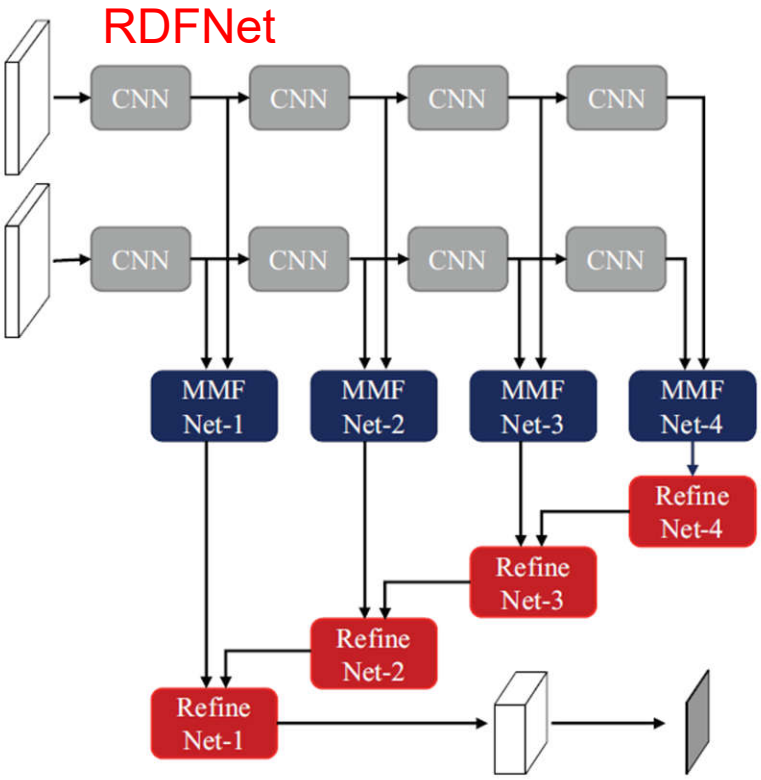
LSTM-CF



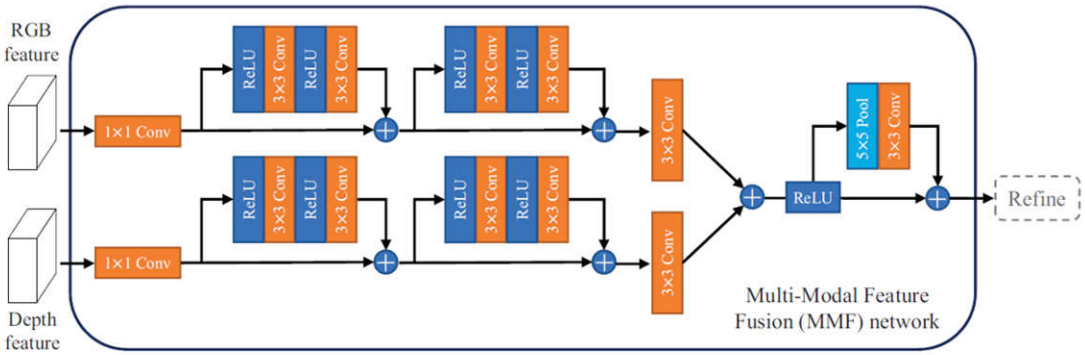
对RGB和D通道分别进行卷积后，利用垂直LSTM进行信息编码，再用水平LSTM进行融合。

[3] Li, Zhen, et al. "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling." *European conference on computer vision*. Springer, Cham, 2016.

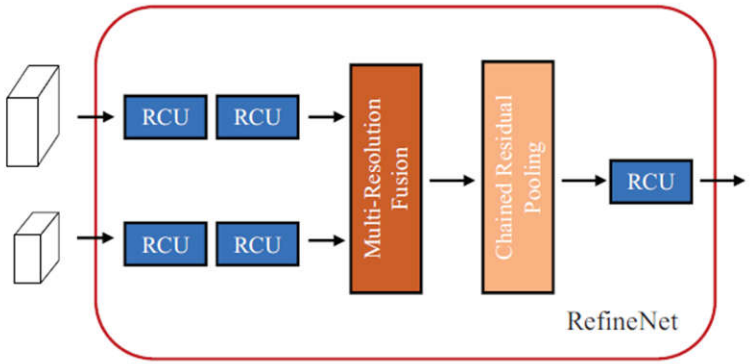
相关工作 RGBD语义分割



(a) 整体流程框图



(b) MMF模块结构图



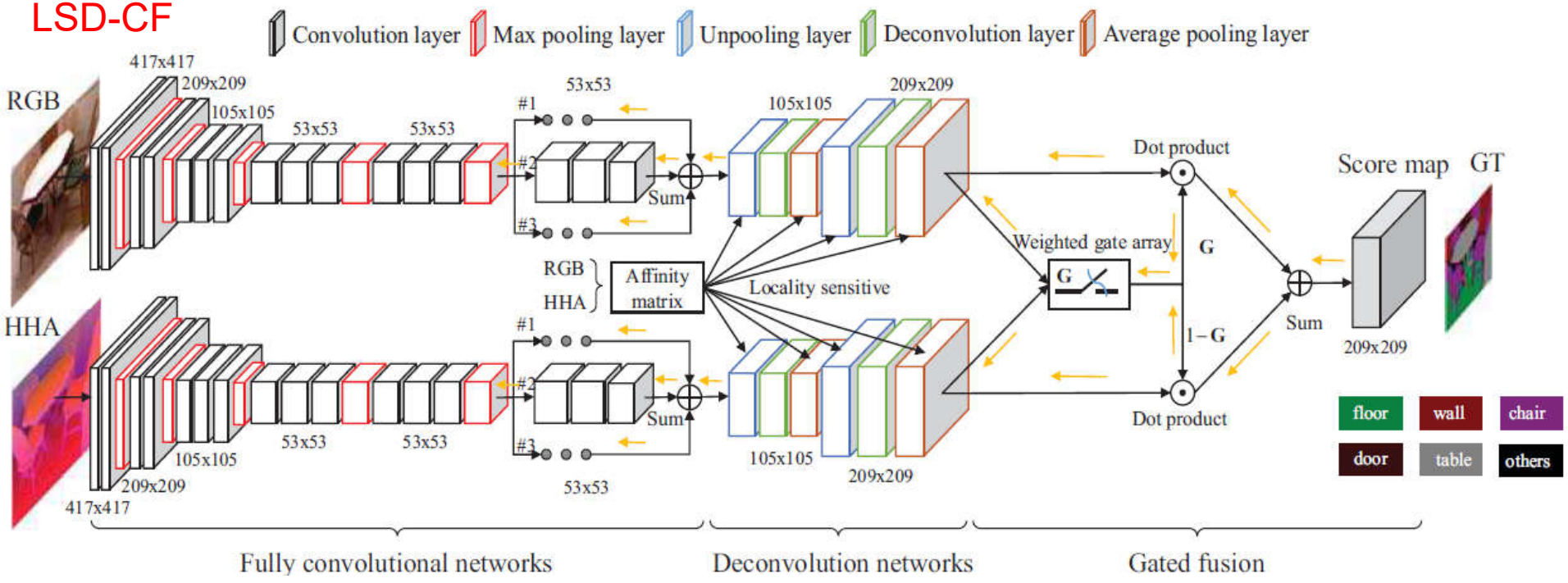
(c) Refine模块结构图

RDFNet扩展了RefineNet提出的RGB语义分割网络，用双通道ResNet作特征提取部分，提出了MMF模块作多模态特征融合。

[4] Park, Seong-Jin, Ki-Sang Hong, and Seungyong Lee. "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

相关工作 RGBD语义分割

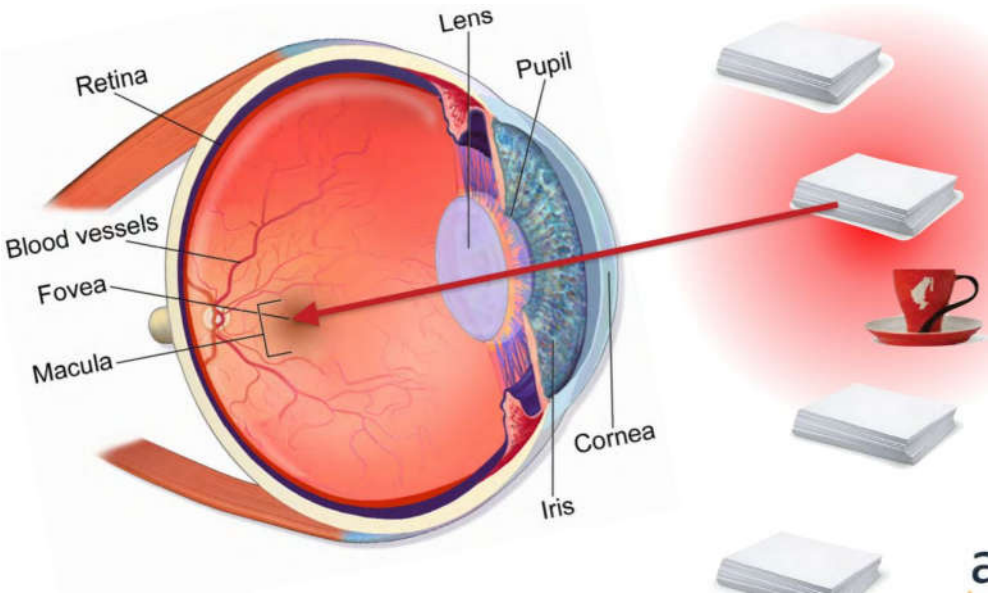
LSD-CF

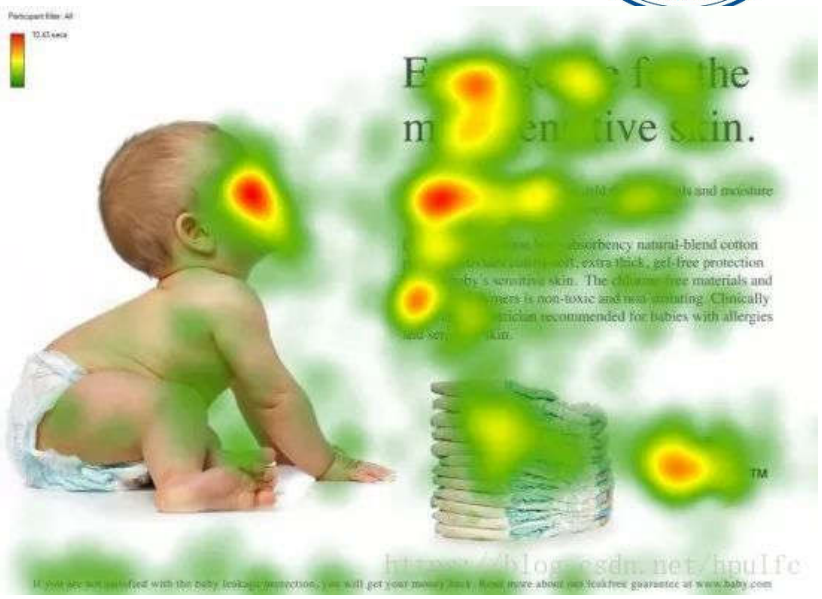


（1）利用多尺度的atros算法来降低分辨率损失，学习鲁棒特征；（2）将邻域RGB-D像素间的成对关系（计算得到的亲和矩阵）嵌入到上采样层和平均池化层中，以恢复特征映射的清晰边界；（3）将RGB和深度评分图合并，学习加权门阵列，以加权每个模态对场景中目标识别的贡献。

[7] Cheng, Yanhua, et al. "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

相关工作 注意力机制



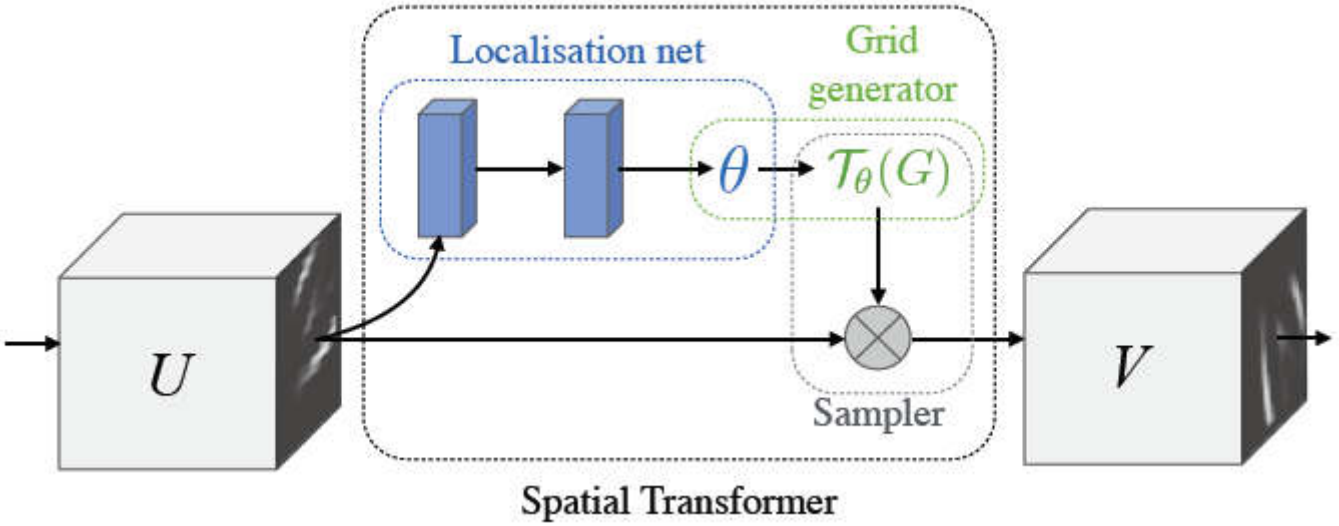


- 注意力机制的核心为聚焦，即关注我们所关心、所需要的部分，在计算机科学领域，我们用加权的方式表征注意力机制。

- 就注意力机制的类型划分，可分为硬注意力（Hard Attention）（不可微分）、软注意力（Soft attention）（可微分）。

- 就注意力机制关注的域来划分，可以分为空间域注意力（Spatial domain attention）、通道域注意力（Channel domain attention）和混合域注意力（Mixed domain attention）。

相关工作 注意力机制

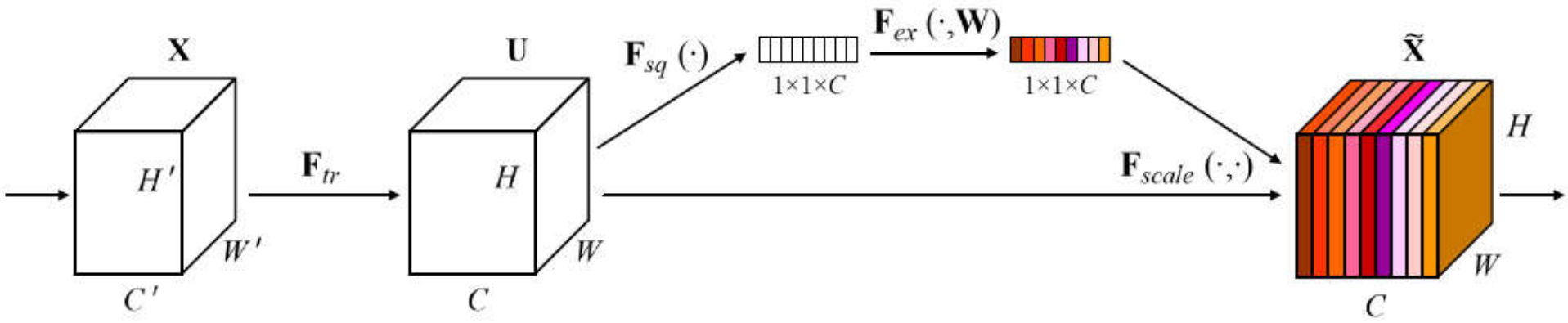


空间转换网络（STN）本质是一种空间域注意力，主要由三部分组成：

1. 本地网络（Localisation network）：由卷积操作或全连接操作组成，得到空间仿射变换参数。
2. 网格生成器（Grid Generator）：根据仿射变换参数构建一个采样网格，得到由输入数据经过采样变换后的输出。
3. 采样器（Sampler）：利用输入的特征图和采样网络的输出特征图进行采样，得到经过空间转换后的结果。

[9] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In *Advances in neural information processing systems* (pp. 2017-2025).

相关工作 注意力机制

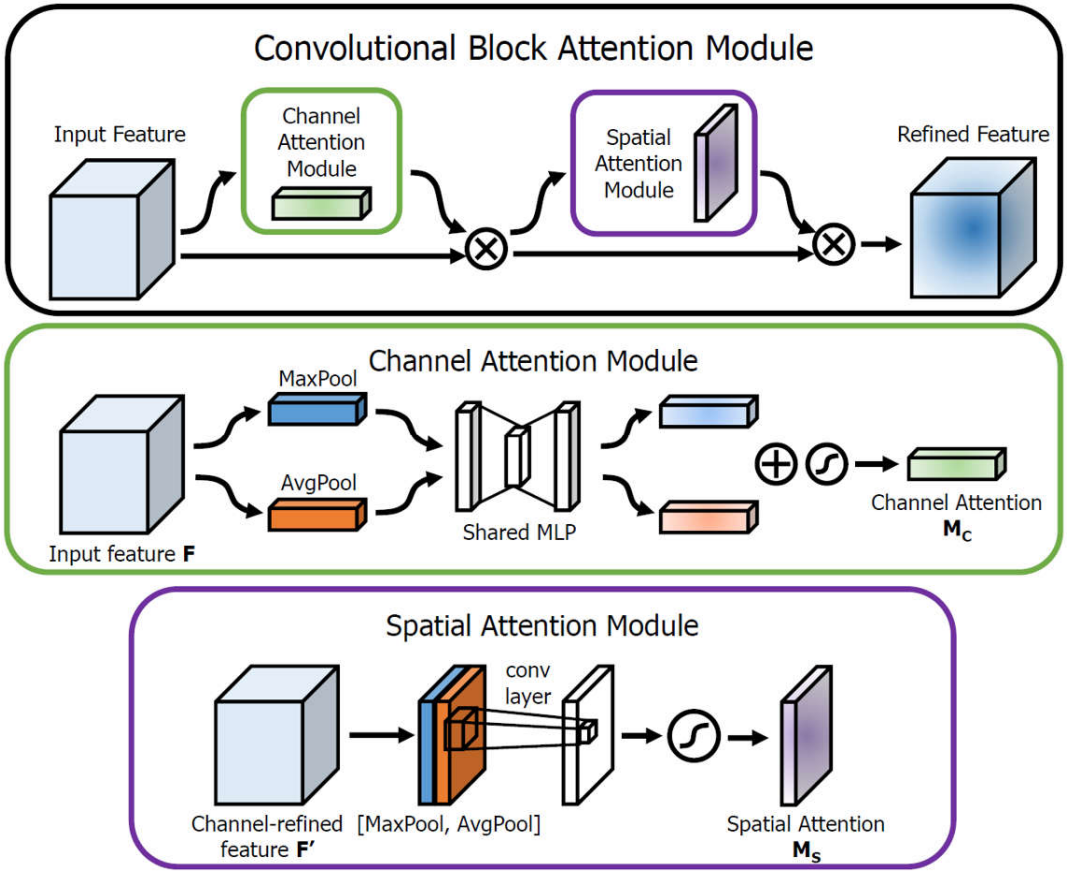


SENet（Squeeze and Excitation Networks）本质是通道域注意力，分为三部分：

1. 挤压（Squeeze）：用全局平均池化（Global Average Pooling）将每个通道内所有特征值相加再平均，得到一个长度与输入特征图通道数一样的特征向量；
2. 激励（Excitation）：利用Bottleneck的结构交换特征，先压缩通道数，再重构通道数，最后利用Sigmod激活函数得到0-1之间的注意力权重；
3. 缩放（Scale）：将得到的通道注意力权重乘回原始输入特征。

[10] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

相关工作 注意力机制



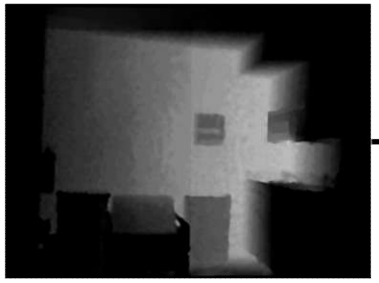
- CBAM——混合域注意力：通道域注意力同时采用最大池化和平均池化操作，再经过Bottleneck的全连接层得到变换结果，然后分别应用于两个通道，最后使用Sigmod激活函数得到通道域的注意力结果。
- 空间域注意力首先将通道本身进行降维，分别获取最大池化和平均池化的结果，然后将其拼接成一个特征图，进过卷积操作融合特征，最后经过Sigmod激活函数得到空间域的注意力结果。

[11] Woo, S., Park, J., Lee, J. Y., & So Kweon, I. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3-19).

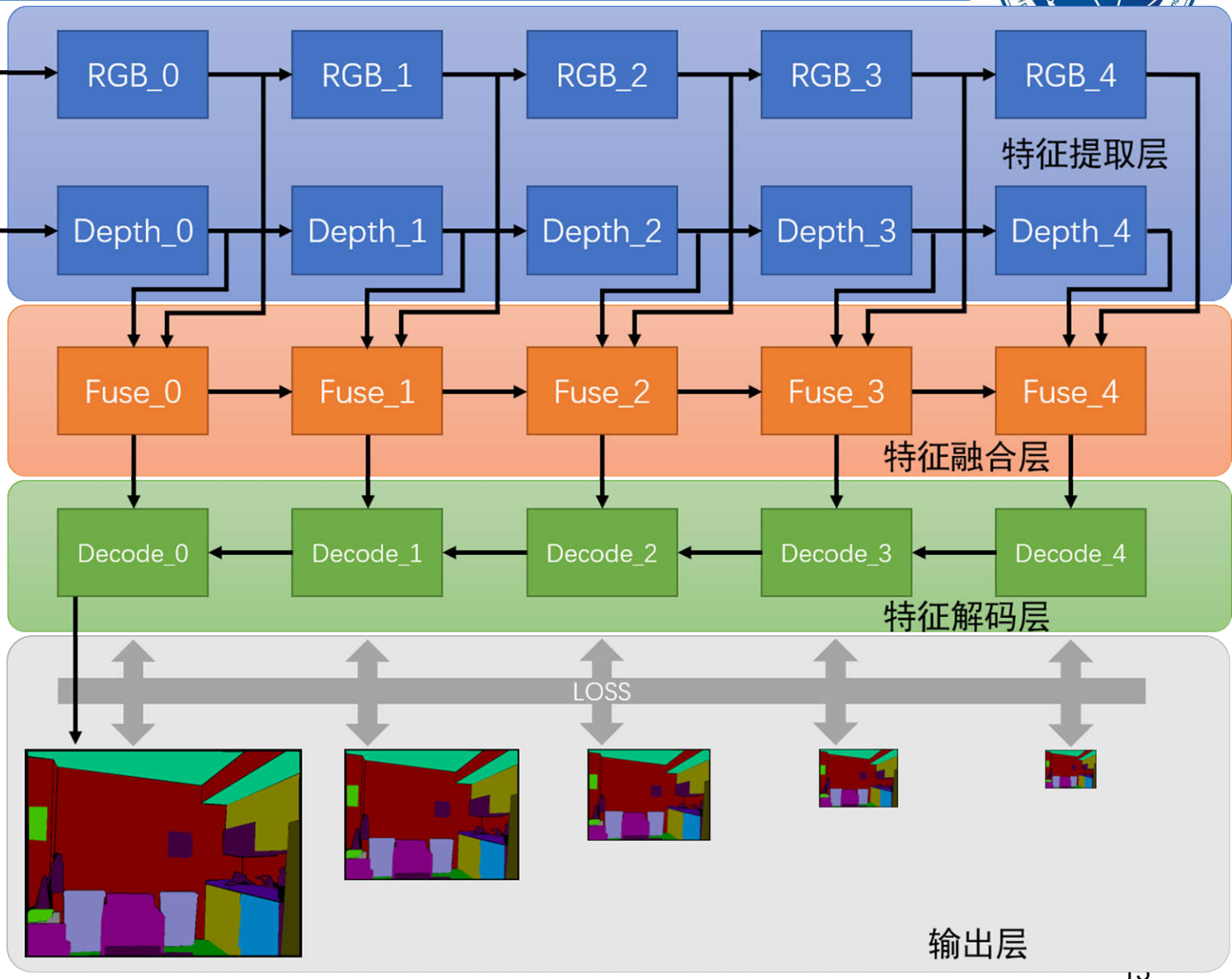
算法设计

整体流程框图

RGB图像



深度图像

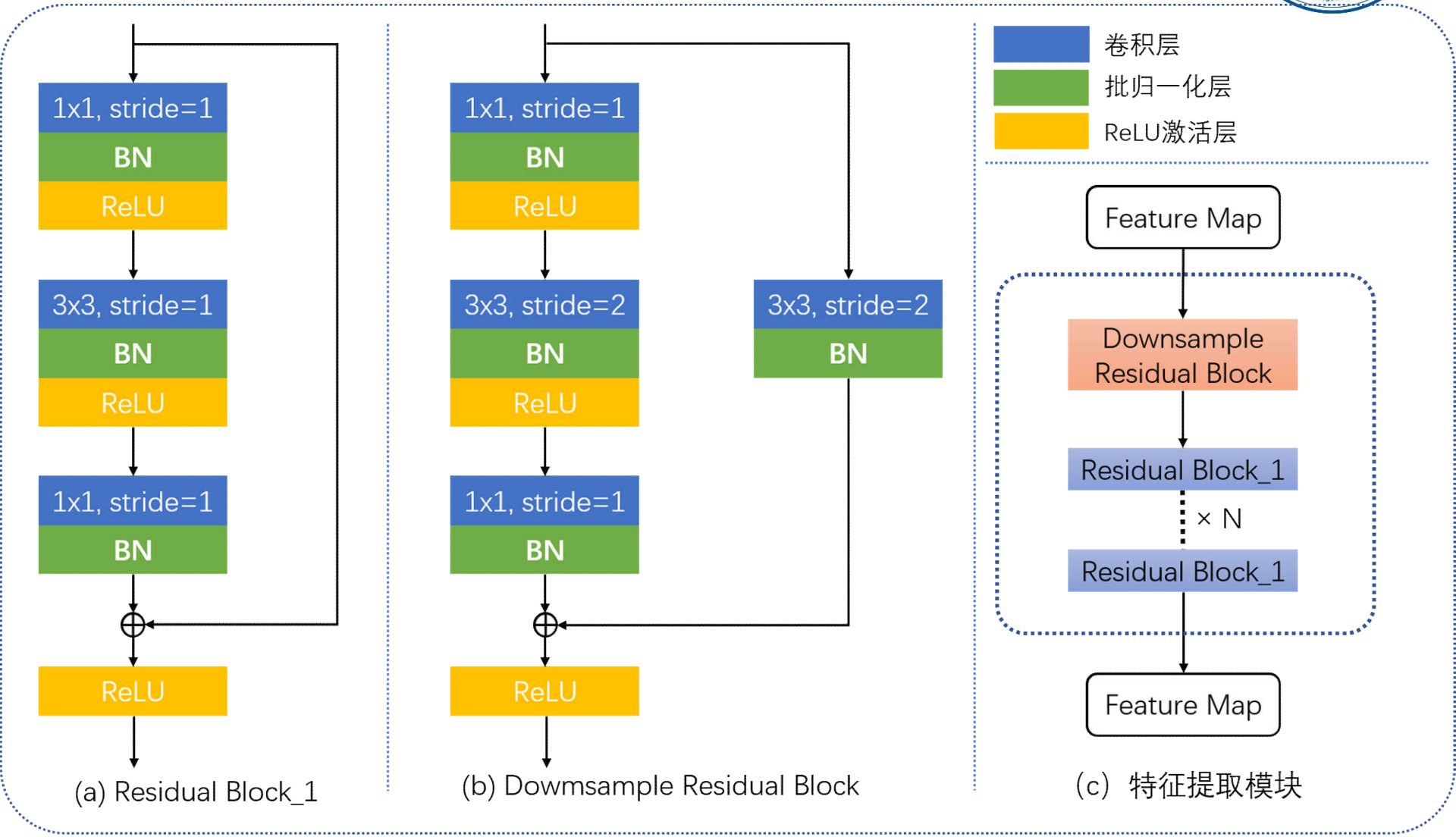


特征提取模块

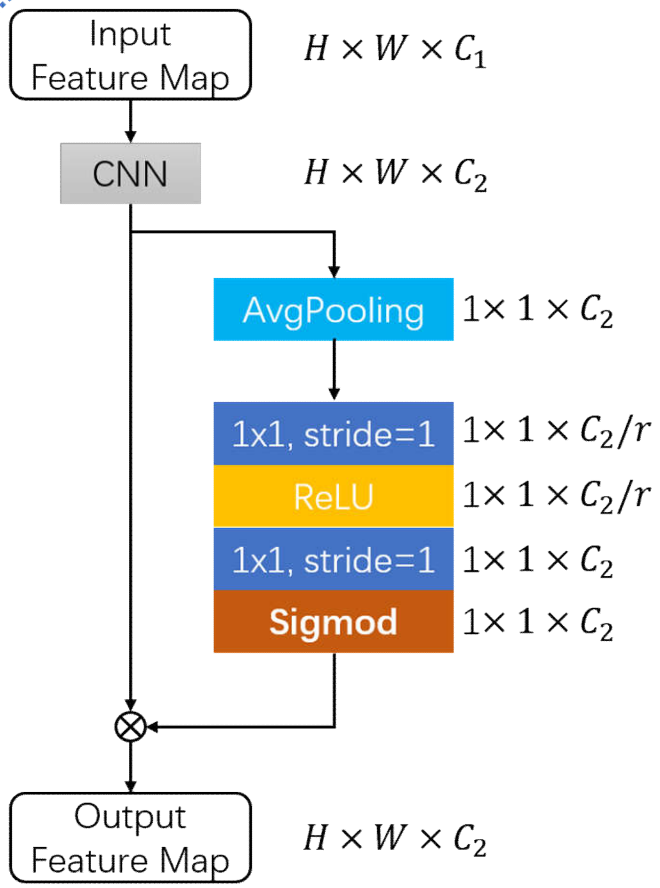
特征融合模块

特征解码模块

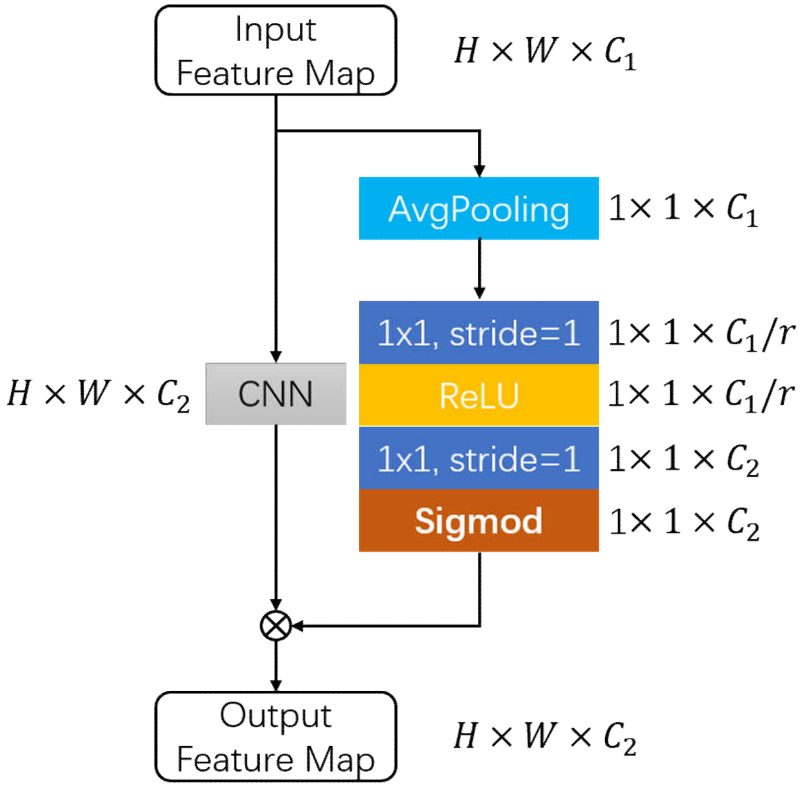
算法设计 特征提取模块设计



算法设计 注意力模块设计

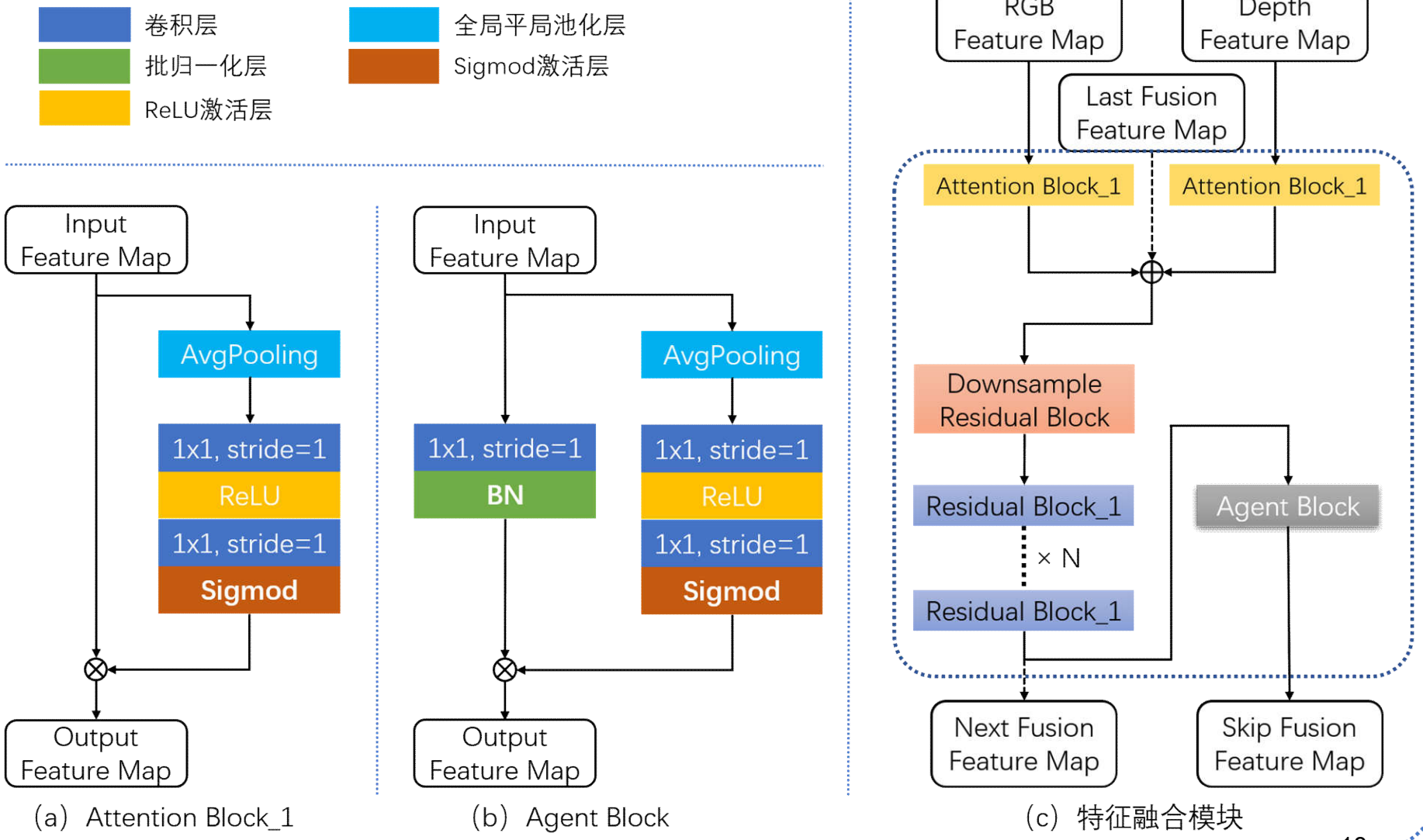


(a) 注意力模块一



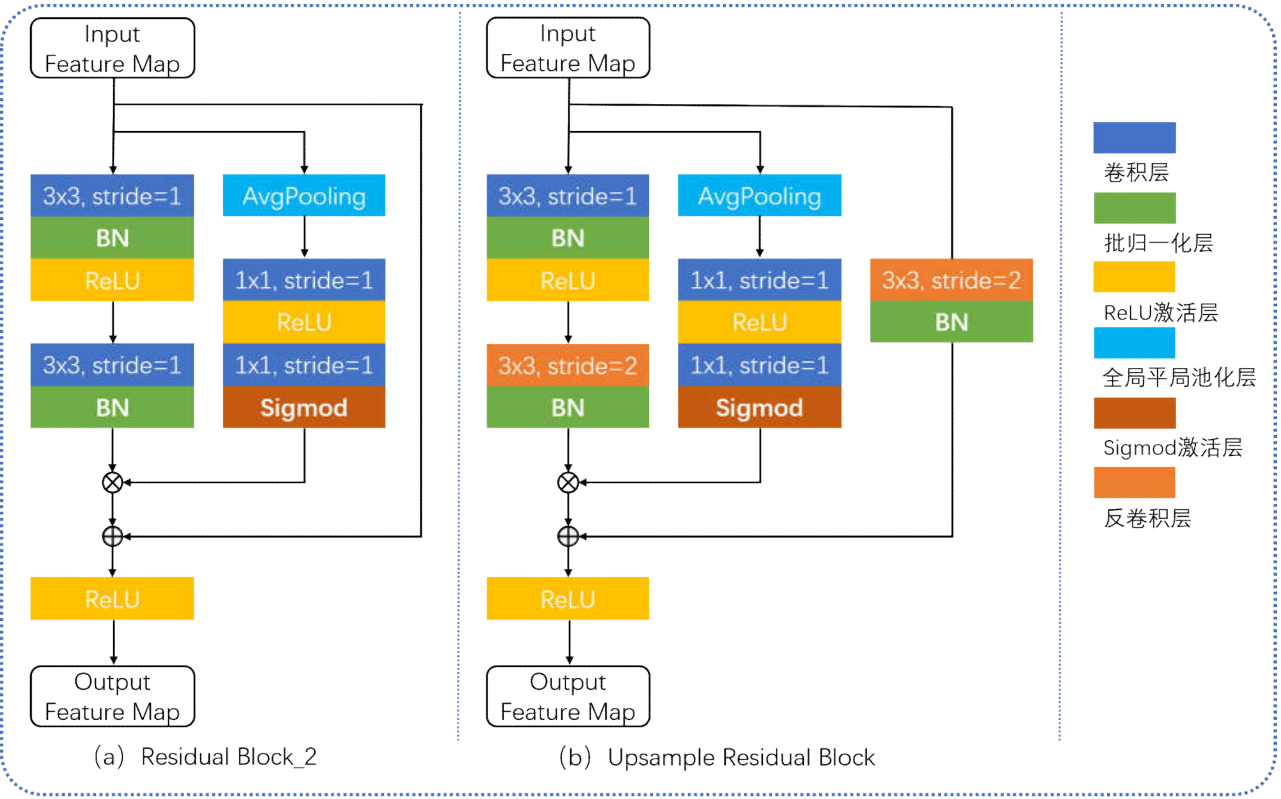
(b) 注意力模块二

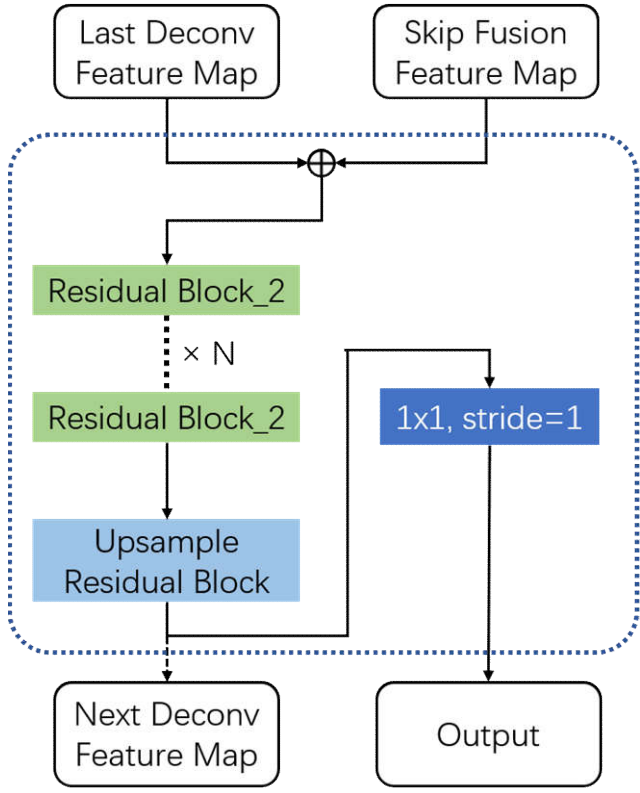
算法设计 特征融合模块设计



算法设计

特征解码模块设计





算法设计 损失函数设计

为解决类别不均衡以及难度不均衡问题，本文采用带权重的Focal Loss，损失函数的计算公式如下：

$$L_k = -\sum_{l}\sum_{c} W_c \times (1 - p_{i,c})^2 \times l^* \times \log(p_{i,c})$$

其中i表示像素，$c \in 1,2,3,\ldots$表示标签类别，$p_{i,c}$表示预测像素i属于类别c的概率，l^*是标签的真实值，W_c为计算类别c损失时的权重。

$$W_c = median(Freq_c)/Freq_c$$

其中$Freq_c$表示每一个类别c出现的频率，$median(Freq_c)$为$Freq_c$的中位数。

[15] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the *IEEE international conference on computer vision* (pp. 2980-2988).

实验结果 数据集

NYU-V2:

- 由微软Kinect摄像头以640*480的分辨率拍摄而成；
- 包含646个不同场景和26中不同场景类型的图像，是目前最流行的室内RGBD数据集；
- 总共包含1449张RGB和深度图像，分割成795张训练集、654张测试集；
- 包含40个类标签。

实验结果 评价指标

n_{ij}：表示标签为i，预测为j的像素个数；

C：表示标签总类别数；

$t_i = \sum_{j=1}^{C} n_{ij}$：表示属于类别i的总像素个数；

Global Pixel Accuracy：$Pixel_{acc} = \frac{\sum_{i=1}^{C} n_{ii}}{\sum_{i=1}^{C} t_i}$；

Mean Pixel Accuracy：$Class_{acc} = \frac{1}{C}\sum_{i=1}^{C} \frac{n_{ii}}{t_i}$；

Mean intersection-over-union (IoU)：$MIoU = \frac{1}{C}\sum_{i=1}^{C} \frac{n_{ii}}{t_i - n_{ii} + \sum_{j=1}^{C} n_{ji}}$；

Weighted intersection-over-union (IoU)：$WIoU = \frac{1}{\sum_{i=1}^{C} t_i}\sum_{i=1}^{C} \frac{t_i * n_{ii}}{t_i - n_{ii} + \sum_{j=1}^{C} n_{ji}}$。

实验结果

实验名称	编码网络	深度图像	预训练模型	跳连结构	金字塔输出	注意力机制	$Pixel_{acc}$	$Class_{acc}$	$MIoU$	$WIoU$
RGB	Resnet50						0.5277	0.3845	0.2188	0.3928
RGBD_cat	Resnet50	√					0.5821	0.4421	0.2723	0.4439
RGBD_split	Resnet50	√					0.6027	0.4690	0.2979	0.4667
RGBD_pretrained	Resnet50	√	√				0.6976	0.6215	0.4080	0.5714
RGBD_skip	Resnet50	√	√	√			0.7101	0.6036	0.4206	0.5866
RGBD_pyramid	Resnet50	√	√	√	√		0.7272	0.6480	0.4677	0.6042
RGBD_attention	Resnet50	√	√	√	√	√	0.7364	0.6609	0.4821	0.6142
RGBD_resnet101	Resnet101	√	√	√	√	√	0.7461	0.6719	0.4918	0.6256
RGBD_resnet152	Resnet152	√	√	√	√	√	**0.7529**	**0.6723**	**0.5013**	**0.6348**

本文方法的组内实验在NYU-V2数据集上的对比结果

实验结果

方法	$Pixel_{acc}$	$Class_{acc}$	$MIoU$
Ren et al.[84]	0.493	0.211	0.214
Gupta et al.[58]	0.591	0.284	0.291
FCN[27]	0.654	0.461	0.340
Liu et al.[49]	0.703	0.517	0.412
LSTM-CF[45]	-	-	0.494
3D Graph[52]	-	0.557	0.431
D-CNN[85]	-	0.563	0.439
Cheng et al.[50]	0.719	0.600	0.459
Lin et al.[86]	-	-	0.477
RDFNet[46]	**0.760**	0.628	0.501
Propose Networks	0.7529	**0.6723**	**0.5013**

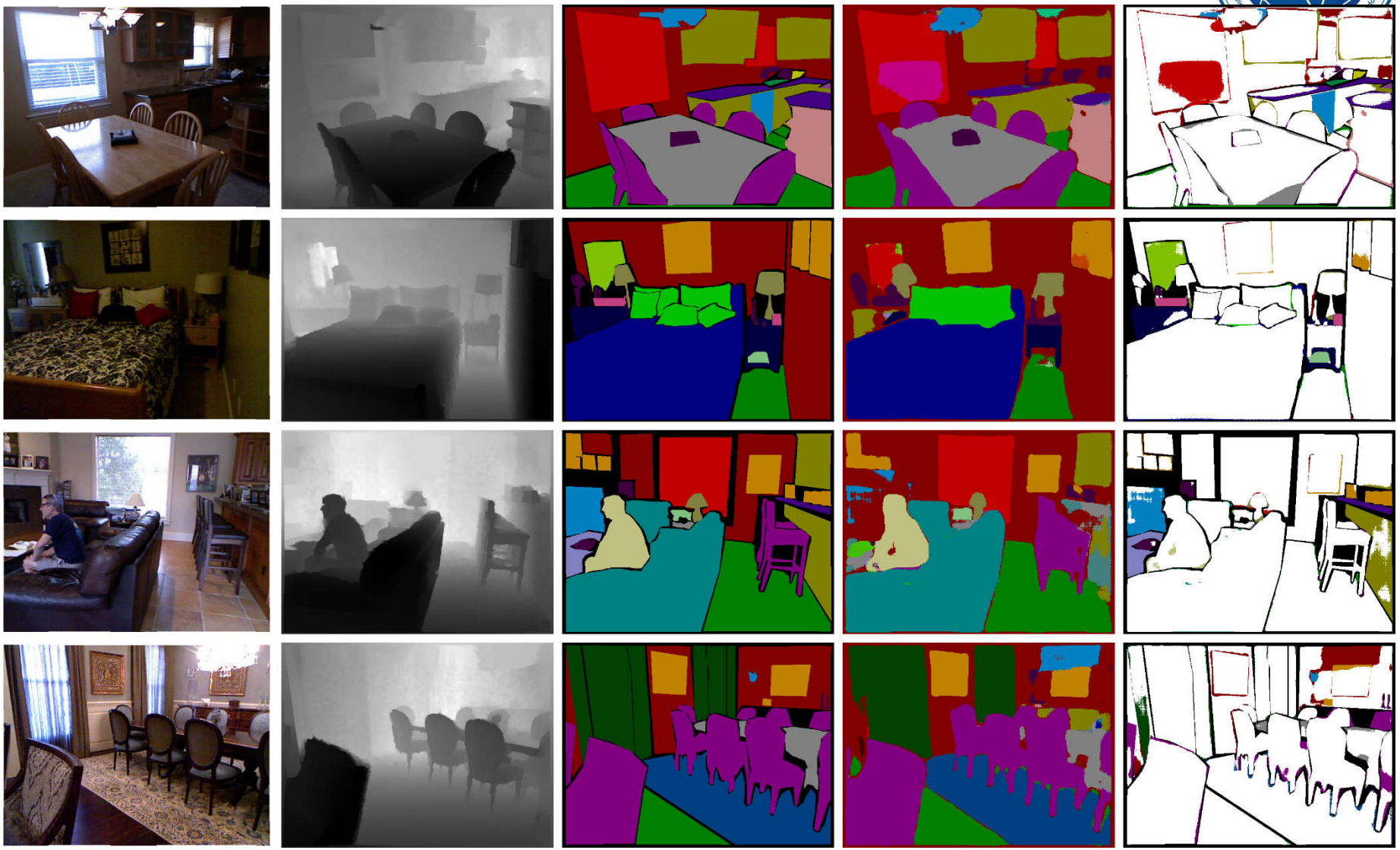
本文方法与其他先进方法在NYU-V2数据集上的对比结果

实验结果

类别	wall	floor	cabinet	bed	chair	sofa	table	door
IoU	0.744	0.8857	0.615	0.7014	0.6386	0.6331	0.4508	0.4285
类别	window	bookshelf	picture	counter	blinds	desk	shelves	curtain
IoU	0.4649	0.4559	0.626	0.6881	0.6092	0.2146	0.1929	0.6548
类别	dresser	pillow	mirror	floor mat	clothes	ceiling	books	refrigerator
IoU	0.5321	0.4691	0.5001	0.4673	0.2362	0.7614	0.3285	0.5427
类别	television	paper	towel	shower curtain	box	whiteboard	person	night stand
IoU	0.6154	0.3418	0.3853	0.4517	0.132	0.7353	0.8158	0.4289
类别	toilet	sink	lamp	bathtub	bag	other structure	other furniture	other prop
IoU	0.7226	0.5896	0.5063	0.4993	0.0793	0.318	0.2013	0.3891

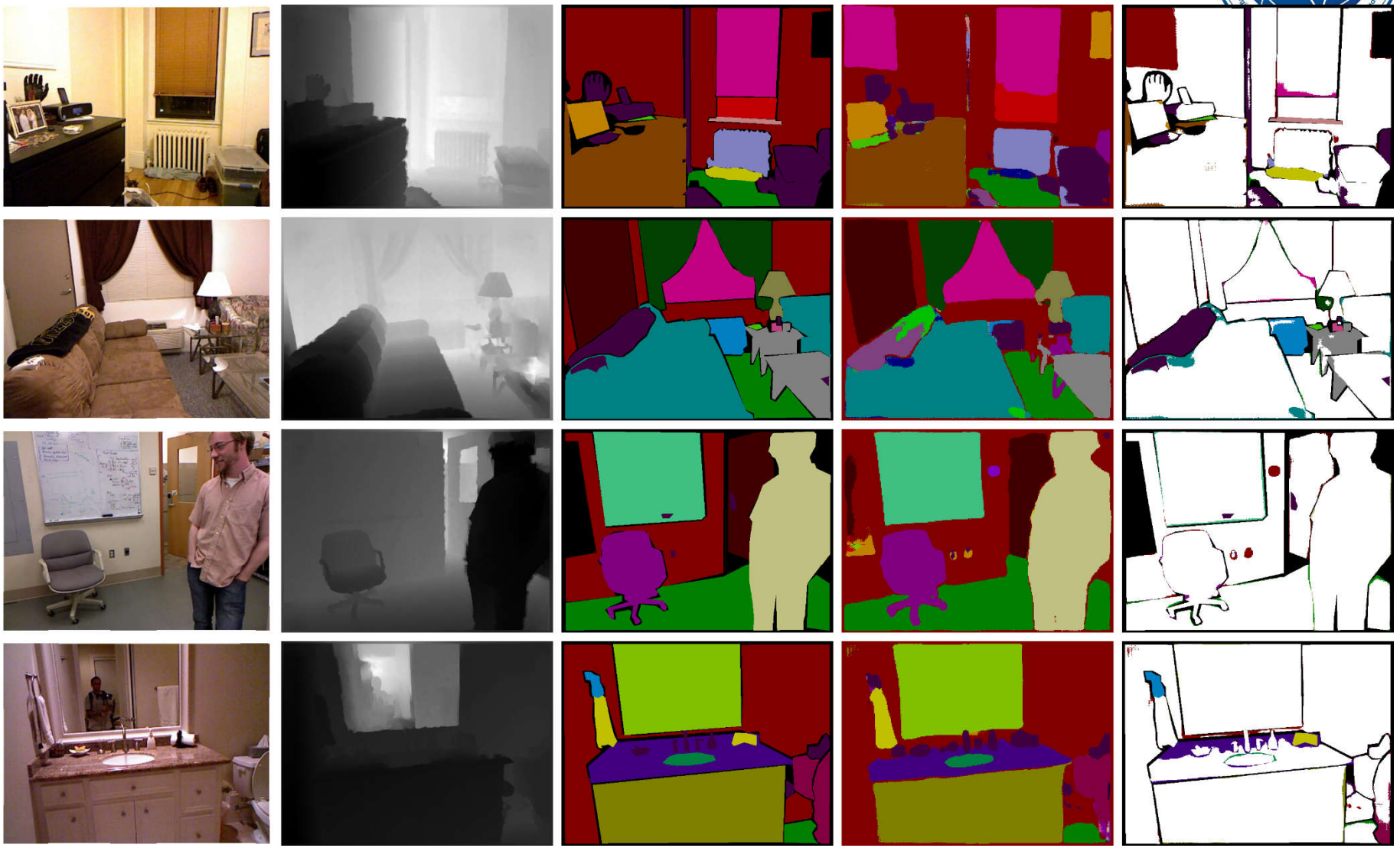
本文方法在NYU-V2数据上每一个类别的分类准确率

实验结果



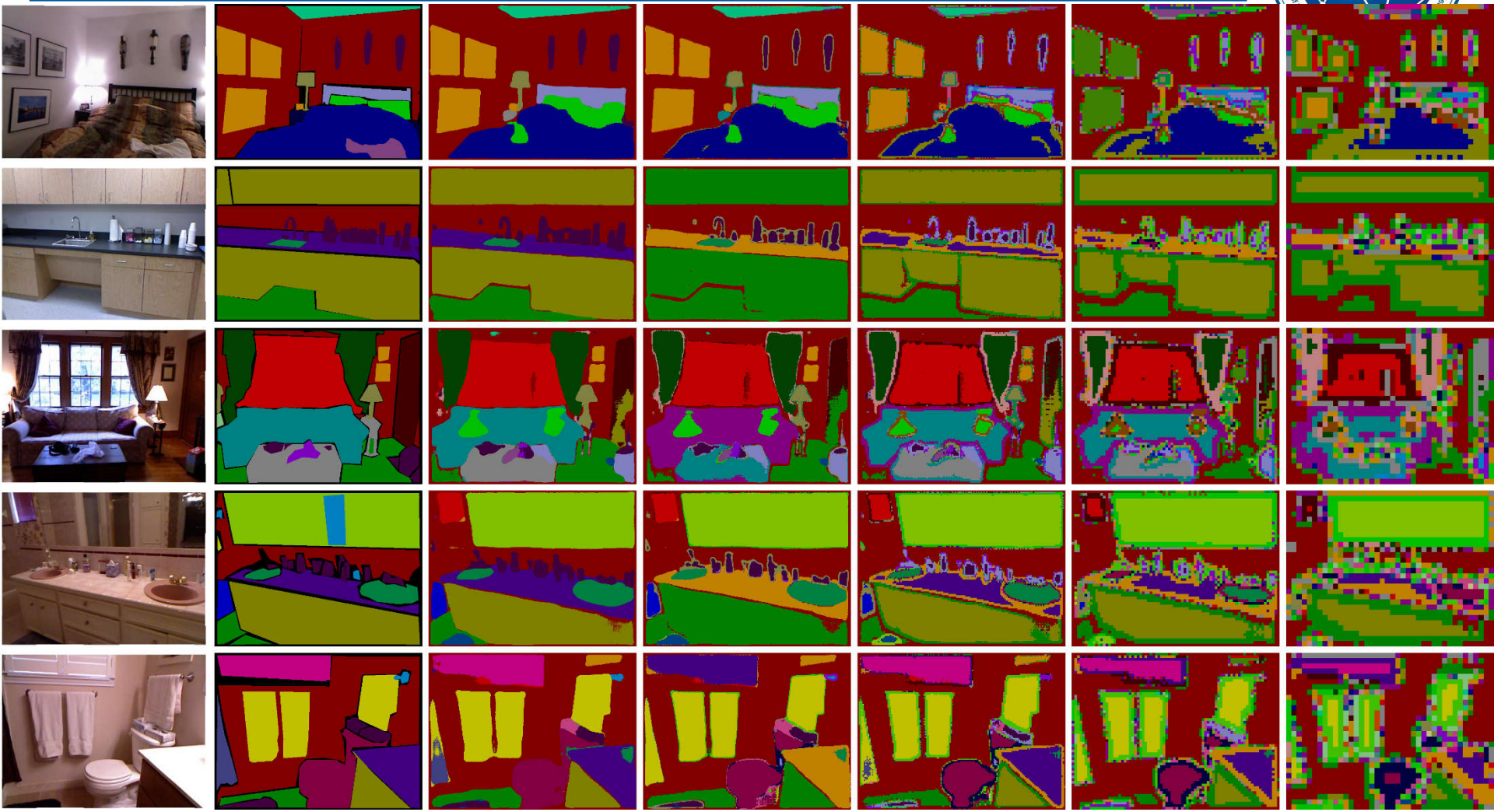
RGB图像　　深度图像　　标记图像　　预测图像　　分析图像

实验结果



RGB图像　　深度图像　　标记图像　　预测图像　　分析图像

实验结果　金字塔层级输出结果



RGB图像　标记图像　Decode_0 输出结果　Decode_1 输出结果　Decode_2 输出结果　Decode_3 输出结果　Decode_4 输出结果

总结

1. 提出了基于编码器解码器架构的注意力融合网络（AFNet），实现了端到端的RGBD图像语义分割。
2. 在RGB和深度图像融合上，提出了单独的特征融合网络，让RGB通路和深度通道分别卷积、互不影响，同时在特征融合模块中引入了注意力机制，实现了两种不同特征图的有效融合；
3. 在跳连结构中引入注意力机制，降低了通道数，减少了计算量，同时保留了足够的空间信息；
4. 改进了Resnet中的残差卷积模块，利用注意力机制增加其在解码器中的全局信息捕获能力；
5. 引入金字塔层级输出，实现了多尺度输出，提高了分割精度；
6. 针对类别不均衡以及难易不均衡现象，引入了带权重的Focal loss，改善了分割结果。

参考文献

[1] Hazirbas, Caner, et al. "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture." *Asian conference on computer vision*. Springer, Cham, 2016.
[2] Wang, Jinghua, et al. "Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks." *European Conference on Computer Vision*. Springer, Cham, 2016.
[3] Li, Zhen, et al. "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling." *European conference on computer vision*. Springer, Cham, 2016.
[4] Park, Seong-Jin, Ki-Sang Hong, and Seungyong Lee. "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
[5] Liu, Hong, et al. "RGB-D joint modelling with scene geometric information for indoor semantic segmentation." *Multimedia Tools and Applications* 77.17 (2018): 22475-22488.
[6] Jiang, Jindong, et al. "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation." arXiv preprint arXiv:1806.01054 (2018).
[7] Cheng, Yanhua, et al. "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
[8] Jiao, Jianbo, et al. "Geometry-aware distillation for indoor semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

参考文献

[9] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In *Advances in neural information processing systems* (pp. 2017-2025).

[10] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

[11] Woo, S., Park, J., Lee, J. Y., & So Kweon, I. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3-19).

[12] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794-7803).

[13] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3146-3154).

[14] Cao, Yue, et al. "Gcnet: Non-local networks meet squeeze-excitation networks and beyond." *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019.

[15] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the *IEEE international conference on computer vision* (pp. 2980-2988).

获奖情况

- 2019年第十五届“挑战杯”广东大学生课外学术科技作品竞赛三等奖

论文

- Automated Steel Bar Counting and Center Localization with Convolutional Neural Networks. *IEEE Transcations on Industrial Informatics*, Under Review. https://arxiv.org/abs/1906.00891
- Retinal Vessel Segmentation via Octave Convolution Neural Network. *IEEE Transcations on Medical Imaging*, Under Review. http://arxiv.org/abs/1906.12193
- Design and Implementation of Mobile Manipulator System. IEEE-CYBER 2019. （已接收）
- Object Sorting Using a Global Texture-Shape 3D Feature Descriptor. *International Journal of Advanced Robotic Systems*, Under Review. https://arxiv.org/abs/1802.01116

专利

- 一种基于深度卷积神经网络的钢筋端面识别方法，专利申请号：201811618063.8
- 一种基于卷积神经网络的电厂电表字符定位和识别方法，专利申请号：201910316734.3
- 一种基于深度学习的智能抓取系统，专利申请号：201810801897.6
- 一种复合型移动机器人（发明），专利申请号：201810780569.2
- 一种复合型移动机器人（实用新型），专利申请号：201821125302.1
- 一种复合型移动机器人及复合型移动机器人控制系统（发明），专利申请号：201810777333.3
- 一种复合型移动机器人及复合型移动机器人控制系统（实用新型），专利申请号：201821125759.2
- 一种构建三维地图的方法，专利申请号：201810809721.5
- 一种基于二维码的移动机器人导航方法，专利申请号：201810809736.1

Thanks!