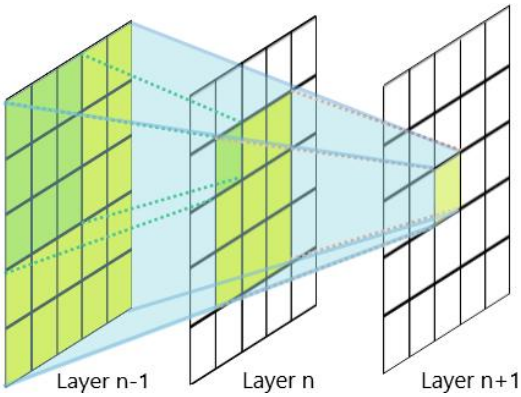
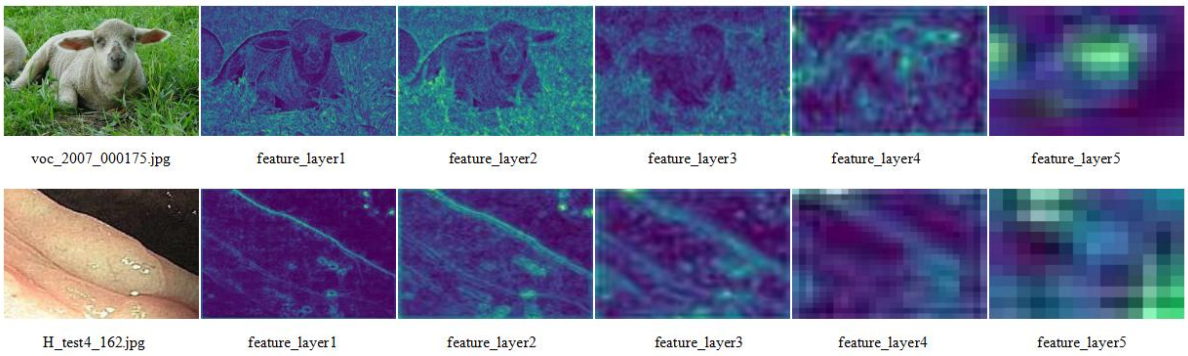
补充内容一（特征分割——跳跃连接）

由于网络层次的深浅和过滤器大小的不同，CNN 可以捕捉低级、高级的特征。其中网络浅层提取的特征与输入比较近，包含更多的像素点的信息。由于网络层数较少所以其特征感受野较小，感受野重叠区域较小，提取到的是图像的细粒度信息，例如图像的一些颜色、纹理、边缘等。因此浅层特征包含更多位置、细节信息，但是由于经过的卷积更少，其语义性更低，噪声更多。网络深层提取的特征离输出较近，包含更多的是更抽象的信息，即语义信息。此外随着网络层数增加图像信息进行压缩、特征感受野不断增大，感受野之间重叠区域增加，提取到的是图像的粗粒度信息，例如图像整体性的一些信息。因此深层特征包含更强的语义信息，但是分辨率很低，对细节的感知能力较差（这种特性在自然图像中更为显著）。



感受野示意图



网络不同层特征提取可视化图

分割是一种精细的分类，需要对图像中的每个像素进行分类。一方面医学分割目标在人体图像中的分布很具有规律，语义简单明确，低分辨率信息能够提供这一信息，用于目标物体的识别。另一方面息肉图像有其特殊性目标（息肉）与其背景（肠道）颜色十分相似、边缘轮廓相比自然图像较为模糊、梯度复杂，故对于息肉分割而言，在分割需要较多的高分辨率信息。

UNet 的编码解码结构恰好能结合低分辨率信息（提供物体类别识别依据）和高分辨率信息（提供精准分割定位依据），完美适用于医学图像分割。因此，在特征分割部分我们借鉴 Unet 的设计结构，通过跳跃连接将高低分辨率特征结合；同时设计与编码部分对称的解码结构逐步恢复目标的细节和相应的空间维度。

补充内容二（特征分类——由 resnet101 换成 vgg19 并且不带 bn）

在本公开数据集的论文中，作者 Krushi PatelI 等人对经典的分类模型：VGG、ResNet、DenseNet、SENet、MnasNet 进行了对比实验。实验结果表明 VGG19 在本数据集上的总体准确率达到 79.78%，优于其他所有模型，这表明在 VGG 之后提出的模型，如 ResNet、SENet 和 MnasNet，虽然在通用图像分类数据集上比 VGG-19 有更好的性能，但在本结肠息肉数据集上都表现不佳。

此外，在结果中还观察到 VGG-19 在大多数指标上都优于带批归一化的 VGG-19，其原因可能是在息肉分类中，像素的精确强度值可能比一般图像分类更有助于区分不同类型的息肉。而批归一化层相对于批缩放像素值，这可能会影响强度信息并降低性能。

Table 3. Evaluation results.

Model	TP	TN	FP	FN	Ade (%)	Hyper (%)	Acc (%)	Err (%)	Pre-1 (%)	Pre-2 (%)	F1-1 (%)	F1-2 (%)	AUC (%)
VGG-19(set-1)	2424	1149	680	466	**83.87**	62.82	**75.71**	**24.28**	78.09	**71.14**	**80.88**	66.72	76.43
VGG-19(set-2)	2419	1346	483	471	**83.70**	**73.59**	**79.78**	**20.21**	**83.35**	**74.07**	**83.52**	**73.83**	84.80
VGG19-BN(set-1)	2071	1440	389	819	71.66	**78.73**	74.40	25.59	**84.18**	63.74	77.42	**70.45**	78.58
VGG19-BN(set-2)	2295	1345	484	595	79.41	73.53	77.13	22.86	82.58	69.32	80.96	71.37	82.20
ResNet50(set-1)	2350	1222	607	540	81.31	66.81	75.69	24.30	79.47	69.35	80.38	68.05	77.25
ResNet50(set-2)	2042	1305	524	848	70.65	71.35	70.92	29.07	79.57	60.61	74.85	65.54	76.27
DenseNet(set-1)	2246	1282	547	644	77.71	70.09	74.76	25.23	80.41	66.56	79.042	68.28	79.28
DenseNet(set-2)	2065	1306	523	825	71.45	71.40	71.43	28.56	79.79	61.28	75.39	65.95	78.65
SENet(set-1)	2230	1320	509	660	77.16	72.17	75.22	24.77	81.41	66.66	79.23	69.30	72.78
SENet(set-2)	2338	1138	691	552	80.89	62.21	73.65	26.34	77.18	62.21	78.99	64.67	82.05
MnasNet(set-1)	2239	1213	616	651	77.47	66.32	73.15	26.84	78.42	65.07	77.94	65.69	73.32
MnasNet(set-2)	2115	1242	587	775	73.18	67.90	71.13	28.86	78.27	61.57	75.64	64.58	77.11

Overall performance of all model on set-1 and set-2 based on individual frame irrespective of sequence.

不同模型结果对比图

考虑到以上原因，我们将原论文特征分类模块的网络结构换成了不带批归一化的 VGG19。这有利于获得更高的精度、使网络更适合结肠息肉图像。

补充内容三（目标检测——用 WSDDN 替换 OICR）

弱监督目标检测开山之作是 WSDDN，但由于 WSDDN 存在丢失实例（1）、容易把邻近的同一类的多个实例检测成同一个实例（2）、检测框容易只框选出实例目标的显著部分，框不全的不足（3）的问题，OICR 被提出以改进 WSDDN。OICR 使用 WSDDN 作为其基线，并在基线后添加了三个实例分类器细化过程，每个实例分类器细化过程由两个完全连接的层组成，旨在进一步预测每个建议的类别分数。

每个实例分类器细化过程的输出是对其后一个细化过程的监督，使更大的区域可以具有比 WSDDN 更高的分数。

在 SDCN 论文中，作者采用 OICR 技术对目标进行检测。然而我们认为这是不合适的：**首先**，就结肠息肉图像而言，虽然临床中存在一张图像中存在大量不同类别息肉的情况，但据目前我们所知的公开的数据集中还并未涉及。已有公开结肠息肉数据集每张图像中仅有一类息肉且不存在一类多实例的情况，因此我们目前工作不受 WSDDN 这一缺点的影响；**其次**，对于框选出实例目标的显著部分，框不全的问题我们的协作机制将会改善；**再者**，对比 WSDDN，OICR 虽精度有所提升但其网络结构复杂，且 OICR 不同类精度结果表明，OICR 精度的提升在于某些符合这一训练策略类的提升，并不是每个类别的精度都提升，有些类别精度反而下降，而结肠息肉图像并不一定符合这种训练策略，因此在 OICR 性能提升策略对于结肠息肉图像并不适用；**最后**，OICR 需要用到三个实例分类器进行细化，与 WSDDN 相比训练更耗时。并且每个实例分类器细化过程的输出是对其后一个细化过程的监督。所以网络在训练中不仅需要保存当前实例分类器的结果、同时还需要保存上一个实例分类器的结果，这会造成更大数据量存储，需要更强大的硬件支持。

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN-VGG_F [4]	42.9	56.0	32.0	17.6	10.2	61.8	50.2	29.0	3.8	36.2	18.5	31.1	45.8	54.5	10.2	15.4	36.3	45.2	50.1	43.8	34.5
WSDDN-VGG_M [4]	43.6	50.4	32.2	**26.0**	9.8	58.5	50.4	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	37.8	46.9	53.4	47.9	34.9
WSDDN-VGG16 [4]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	**42.6**	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
WSDDN+context [16]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	**49.2**	**42.0**	47.3	56.6	15.3	12.8	24.8	**48.9**	44.4	47.8	36.3
OICR-VGG_M	53.1	57.1	**32.4**	12.3	**15.8**	58.2	56.7	39.6	0.9	**44.8**	39.9	31.0	**54.0**	62.4	4.5	20.6	39.2	38.1	48.9	48.6	37.9
OICR-VGG16	**58.0**	**62.4**	31.1	19.4	13.0	**65.1**	**62.2**	28.4	**24.8**	44.7	30.6	25.3	37.8	**65.5**	**15.7**	**24.1**	**41.7**	46.9	**64.3**	**62.6**	**41.2**
WSDDN-Ens. [4]	46.4	58.3	35.5	**25.9**	14.0	66.7	53.0	39.2	8.9	41.8	26.6	**38.6**	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.3
OM+MIL+FRCNN [20]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	**46.1**	**21.8**	57.1	22.1	34.4	50.5	61.8	**16.2**	**29.9**	40.7	15.9	55.3	40.2	39.5
OICR-Ens.	58.5	63.0	35.1	16.9	17.4	63.2	60.8	34.4	8.2	49.7	41.0	31.3	51.9	64.8	13.6	23.1	41.6	48.4	58.9	58.7	42.0
OICR-Ens.+FRCNN	**65.5**	**67.2**	**47.2**	21.6	**22.1**	**68.0**	**68.5**	35.9	5.7	**63.1**	**49.5**	30.3	**64.7**	**66.1**	13.0	25.6	**50.0**	**57.1**	**60.2**	**59.0**	**47.0**

Table 2. Average precision (in %) for different methods on VOC 2007 test set. The upper part shows results using a single model. The lower part shows results of combing multiple models.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
WSDDN-VGG_F [4]	68.5	67.5	56.7	34.3	32.8	69.9	75.0	45.7	17.1	68.1	30.5	40.6	67.2	82.9	28.8	43.7	71.9	**62.0**	62.8	58.2	54.2
WSDDN-VGG_M [4]	65.1	63.4	**59.7**	45.9	38.5	69.4	77.0	50.7	30.1	68.8	34.0	37.3	61.0	82.9	25.1	42.9	**79.2**	59.4	68.2	64.1	56.1
WSDDN-VGG16 [4]	65.1	58.8	58.5	33.1	**39.8**	68.3	60.2	**59.6**	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
WSDDN+context [16]	**83.3**	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	**47.1**	**59.5**	67.0	83.5	**35.3**	39.9	67.0	49.7	63.5	65.2	55.1
OICR-VGG_M	81.7	72.9	56.5	31.4	36.3	75.6	81.6	57.0	7.3	74.7	**47.1**	46.0	**78.2**	**88.8**	12.2	46.2	66.0	56.7	65.8	64.9	57.3
OICR-VGG16	81.7	**80.4**	48.7	**49.5**	32.8	**81.7**	**85.4**	40.1	**40.6**	**79.5**	35.7	33.7	60.5	**88.8**	21.8	**57.9**	76.3	59.9	**75.3**	**81.4**	**60.6**
OM+MIL+FRCNN [20]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	**55.2**	28.5	68.8	18.5	**49.2**	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
WSDDN-Ens. [4]	68.9	68.7	**65.2**	42.5	40.6	72.6	75.2	53.7	**29.7**	68.1	33.5	45.6	65.9	86.1	**27.5**	44.9	**76.0**	62.4	66.3	66.8	58.0
OICR-Ens.	85.4	78.0	61.6	40.4	38.2	82.2	84.2	46.5	15.2	80.1	45.2	41.9	73.8	89.6	18.9	56.0	74.2	62.1	73.0	77.4	61.2
OICR-Ens.+FRCNN	**85.8**	**82.7**	62.8	**45.2**	**43.5**	**84.8**	**87.0**	46.8	15.7	**82.2**	**51.0**	45.6	**83.7**	**91.2**	22.2	**59.7**	75.3	**65.1**	**76.8**	**78.1**	**64.3**

Table 3. CorLoc (in %) for different methods on VOC 2007 trainval set. The upper part shows results using a single model. The lower part shows results of combing multiple models.

OICR 不同类精度结果

所以，我们改进了 SDCN 的目标检测网络框架以适应结肠息肉图像的检测，同时降低网络架构对硬件的要求，减少训练时间。