Fully Supervised Image Segmentation

Tieqiao Wang*
Oregon State University
1148 Kelley Engineering Center
wangtie@oregonstate.edu

Thomas Noel*
Oregon State University
1148 Kelley Engineering Center
noelt@oregonstate.edu

Ying Dai*
Oregon State University
Kidder Hall, M114
daiy@oregonstate.edu

Abstract

Semantic image segmentation is a critical component in most visual understanding tasks, for example, medical image analysis, autonomous driving and robotic perception. Recently, numerous models based on deep learning have been proposed to solve this problem. Among them, U-net, DeepLabv3+ and OCR are classic ones and achieves new state of the art performance. In this paper, we explore these three methods on two different datasets, i.e., the magnetic resonance imaging (MRI) dataset and PASCAL VOC 2012 benchmark. For the OCR method, we have proposed multiple variants based on the transformation and loss functions and demonstrated the effectiveness of the proposed model on MRI dataset, achieving the validation accuracy of 92.1%(mIOU).

1. Introduction

The applications of image segmentation are everywhere nowadays. For medical image analysis, accurate segmentation of brain magnetic resonance imaging (MRI) is an essential step in quantifying the changes in brain structure.[5] And for autonomous driving, it is obvious that assigning labels correctly is very important for right decision making. Recently, deep learning models have been more and more popular to tackle such problems, giving their promising and impressive performance. However, there are two main challenging of semantic segmentation in general. First, it is extremely hard to collect a large dataset for training, especially for medical image data, due to limited number of patients, but the success of deep learning was limited based on the size of training sets. Additionally, in real life settings, the objects may have complex relationship with each other such as occlusion and some of them are extremely tiny and slim to precisely segment. Motivated by the board applications and these challenges, numerous works have been proposed. U-net [8] is proposed to work with small training size and still yields precise segmentation. It replaces the pooling operators by upsampling operators, which increases the resolution of the output. Besides, U-net combines the high resolution features from the contracting path with the upsampled output to localize. Using the assembled information, U-net can then learn a more precise output. Afterwards, spatial pyramid pooling model and encode-decoder structure are used in deep neural networks for semantic segmentation task. More specifically, spatial pyramid pooling model can encode multi-scale contextual information and encode-decoder structure can capture sharper object boundaries by recovering the spatial information. DeepLabv3+ is then proposed combining advantages from both methods. Inspired by both U-net and DeepLabv3+, we find one latest model which consider using object-contextual representation (OCR) in the neural network. In the OCR model, the relation between each pixel and each object region is computed and it will be used to update the final representation for each pixel. In this project, we use U-net as the baseline and compare the performance with both DeepLabv3+ and OCR model. We implement several versions of OCR model and the one using "1×1 conv $\rightarrow$ BN $\rightarrow$ ReLU" as transformation function and lovasz-softmax loss has the best accuracy which is measured by mean Intersection-Over-Union.

In Section 2 , we will briefly review the milestone innovations of semantic segmentation. In Section 3 , we introduce the architecture of U-net, DeepLabv3+ and OCR pipeline. These three models are implemented and tested on both MRI dataset and PASCAL VOC 2012 benchmark. Additionally, our experiment settings and results are discussed in Section 4. Section 5 and Section 6 present our conclusion and future work.

2. Related Work

2.1. Fully Convolutional Network

Fully convolutional network(FCN) [7] is a pioneering work using deep learning methods for semantic image segmentation. This is a naive solution to transfer classification CNNs to segmentation tasks by replacing fully-connected

*These three authors contributed equally to this work.

layers with fully-convolutional layers. Then, the model combines features from early layers by using skip connections and upsampling. Despite its simple design, it achieved state of the art segmentation performance when it was presented.

2.2. Encoder-Decoder

The encoder-decoder architectures (i.e., SegNet [1], U-net [8]) gradually reduce the feature representation size and capture higher semantic information in the encoder module, and then gradually recover the spatial information in the decoder.

2.3. Multi-scale Context

To capture multi-scale information, there are several interesting works, for example, PSPN [14] and DeepLabv3+ [3]. In PSPN architecture, a CNN produces the feature map and a pyramid pooling module aggregates the different sub-regions. Then, the final feature representation is obtained from up-sampling and concatenation. In the end, the final pixel-wise prediction is generated by applying convolution to the final feature represention. DeepLab family applies atrous convolution with different rates (also called Atrous Spatial Pyramid Pooling, or ASPP) to enhance feature representation at different scales. Some of these models are still considered to be state-of-the-art.

2.4. Relational Context

DANet[4] and OCNet[12] enhance the representation for each pixel by considering the contextual pixels, where the context is computed based on the relation (or similarity) between all the pixels (self-attention scheme). There are also some works, like ACFNet[13], that augment the pixel representations by utilizing the region representations.

3. Technical Approach

3.1. U-net

The network architecture is illustrated in Figure 5. Since the network architecture has a U-shape, it is called U-net. In general, we could divide U-net into two parts, the left side is called a contracting path and the right side is call an expansive path. The contracting path consists of the repeated application of 3×3 unpadded conv $\rightarrow$ ReLU $\rightarrow$ 2×2 max pool with stride 2. At each downsampling step, the number of feature channels is doubled. The expansive path consists of the repeated application of 2×2 conv $\rightarrow$ 3×3 conv $\rightarrow$ ReLU. Note that after upsampling using 2×2 conv , the representation is concatenated with the corresponding cropped feature map from the contracting path, then passes through a typical convolutional network. The last layer is a 1×1 conv and it is used to map each component feature to the desired number of classes.

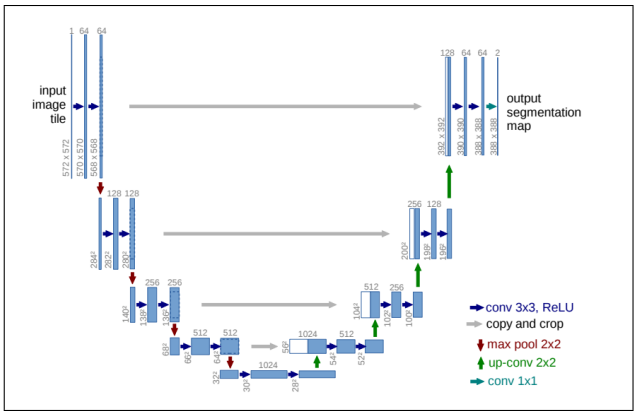

Figure 1. U-net architecture. Each blue box corresponds to multi-channel feature map. The number of channels is denoted on top of the box. They x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

3.2. DeepLabv3+

As shown in Figure 2, the architecture of DeepLabv3+ can be divided to two parts, an encoder and a decoder. The encoder employs atrous convolution to extract the features computed by deep convolutional neural networks. Furthermore, it augments the Atrous Spatial Pyramid Pooling module, which probes convolutional features at multiple scales by applying atrous convolution with different rates, with the image-level features. To better recover object segmentation details, in the decoder part, it concatenates then encoder features (upsample by 4) with the corresponding low-level features from the network backbone.

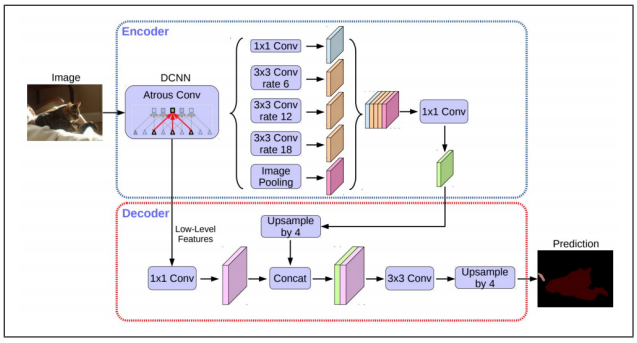

Figure 2. DeepLabv3+ architecture. The encoder module encodes multi-scale contextual information by applying atrous convolution at multiple scales, while the simple yet effective decoder module refines the segmentation results along object boundaries.

3.3. OCR

The pipeline of OCR is shown in Figure 3. In this project, we used HRNet-W48 [9] (with output stride 4) as

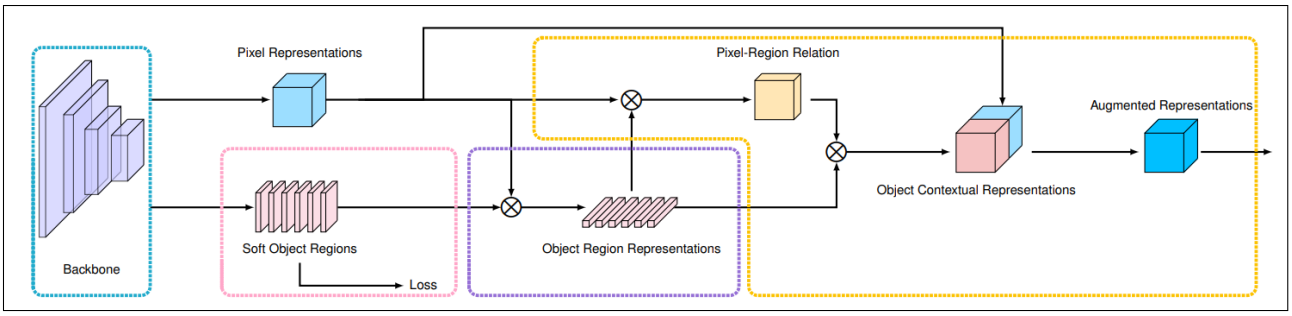

Figure 3. Illustrating the pipeline of OCR

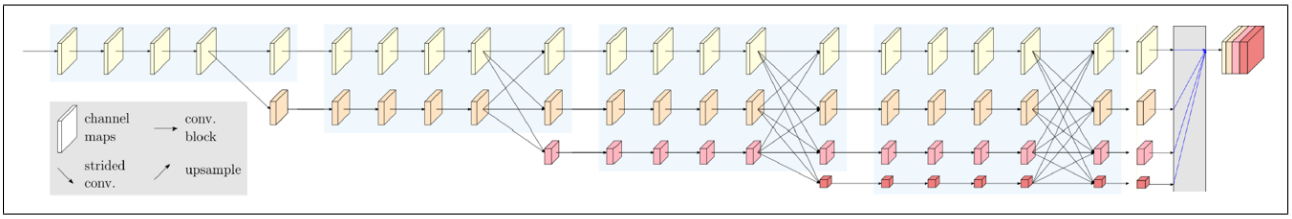

Figure 4. Illustrating the pipeline of OCR

the backbone. The primary benefit of this high-resolution network is that it maintains high-resolution representations through the whole process by connecting high-to-low resolution convolutions in parallel and produces strong high-resolution representations by repeatedly conducting fusions across parallel convolutions. The network architecture for HRNet is presented in Figure 4. It aggregates the output representations at four different resolutions, and then uses a 1×1 convolution to fuse these representations. The output from HRNet will go through a 1×1 convolution layer to predict the coarse segmentation (soft object region) supervised with a pixel-wise cross-entropy loss. After this step, each image is partitioned to K soft object regions$\{\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_K\}$. The object region representation is calculated by the following formula,

$$\boldsymbol{f}_k = \sum_{i \in \mathcal{I}} \tilde{m}_{ki} \boldsymbol{x}_i \tag{1}$$

where $\boldsymbol{x}_i$ is the representation of pixel p_i and $\tilde{m}_{ki}$ is the normalized degree of pixel p_i belonging to the kth object region.

Then, we use both the representation of pixel p_i and the object region representation to compute the relation between each pixel and each object region as below:

$$w_{ik} = \frac{e^{\kappa(\boldsymbol{x}_i, \boldsymbol{f}_k)}}{\sum_{j=1}^{K} e^{\kappa(\boldsymbol{x}_i, \boldsymbol{f}_k)}} \tag{2}$$

Where $\kappa\left(\boldsymbol{x}, \boldsymbol{f}\right) = \phi(\boldsymbol{x})^T \psi(\boldsymbol{f})$ is the unnormalized relation function.

The objective contextual representation $\boldsymbol{y}_i$ for pixel p_i is computed by aggregating the object region representations via consideration of its relations which all the object region, which formulated as

$$\boldsymbol{y}_i = \rho\left(\sum_{k=1}^{K} w_{ik} \delta(\boldsymbol{f}_k)\right) \tag{3}$$

The final representation for pixel p_i is updated as the aggregation of the original representation $\boldsymbol{x}_i$ and the object contextual representation $\boldsymbol{y}_i$,

$$\boldsymbol{z}_i = g\left([\boldsymbol{x}_i^T \boldsymbol{y}_i^T]\right) \tag{4}$$

where $g(\cdot)$ is a transform function used to fuse the original representation and the object contextual representation.

4. Experiments

In this project, we train and implement U-net, DeepLabv3+(ResNet101) and HRNet+OCR using PASCAL VOC. In this case, HRNet+OCR is only a little bit worse than DeepLabv3+. We then focus on modifying the HRNet+OCR model for MRI data. We tried one other activation function and two other loss functions. In total, six models were applied to MRI data with 100 epochs. All the experiments were conducted on NVIDIA Tesla V100 SXM3 with batch size = 8 and learning rate = 0.01.

4.1. Dataset

PASCAL VOC One of the datasets used in our experiments is the PASCAL VOC dataset which contains 20 diverse object classes presented in a variety of settings in addition to one background class. This dataset has been used widely

as a benchmark for the semantic segmentation task. It has 1,464 (train), 1,449 (val), and 1,456 (test) pixel-level annotated images.

MRI The Brain MRI dataset we used in our project was obtained from The Cancer Genome Atlas (TCGA) and The Cancer Imaging Archive (TCIA). There are 110 patients from TCGA lower-grade glioma collection and each of them have around 20-40 images conveying tumor information, each associated with a tumor mask. Ground truth segmentation labels were provided by medical experts. The training set is approximately 700 MB and the test set is around 300 MB.

4.2. Metrics

In semantic segmentation, **Intersection over Union** (IOU) also known as the **Jaccard Index** is the most commonly used metric. It is defined as

$$IOU = J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

where A and B represent the ground truth and the segmentation results.
We evaluate our experimental results based on mIOU, which is the average IOU over all classes.

4.3. Activation Function Ablation

Referring to Eqs. 3 and 4, ρ, δ, and g are all transformation functions described by 1×1 conv $\rightarrow$ BN $\rightarrow$ ReLU. In an attempt to boost system performance an experiment was run using leaky ReLU [10] in lieu of ReLU. Recall that ReLU is described by

$$h(x) = \max\{0, x\} \tag{6}$$

whereas leaky ReLU is simply described by

$$h(x) = \begin{cases} x, & x \geq 0 \\ 0.01x, & \text{otherwise} \end{cases} \tag{7}$$

4.4. Loss Function Ablation

In addition to the baseline cross entropy loss used in [11], experiments were conducted using both the focal loss [6] and the Lovasz-Softmax loss [2] as well. The cross entropy loss is described by

$$\mathrm{CE}(\boldsymbol{f}, k) = -\log\left(\frac{e^{f_k}}{\sum_{i=1}^{N} e^{f_i}}\right) \tag{8}$$

The focal loss is typically used to address class imbalance issues which can be rectified via hyperparameter tuning,

Table 1. PASCAL VOC 2012 val set results.

Method	mIOU(%)
UNet (80 epochs)	37.6
DeepLabv3+ (80 epochs)	79.1
HRNet+OCR (100 epochs)	78.4

Table 2. MRI val set results.

Method	mIOU(%)
UNet	90.9
DeepLabv3+	90.0
HRNet+OCR (baseline)	91.5
HRNet+OCR (leakyReLU)	91.5
HRNet+OCR (focal loss)	91.3
HRNet+OCR (lovsz-softmax loss)	**92.1**

$$\mathrm{FL}(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t) \tag{9}$$

where α_t is the class-balancing parameter and $\gamma \geq 0$ is the focusing parameter. Additionally, we used the Lovasz-Softmax loss

$$\mathrm{loss}(\boldsymbol{f}) = \frac{1}{|\mathcal{C}|}\sum_{c \in \mathcal{C}} \overline{\Delta_{J_c}}(\boldsymbol{m}(c)) \tag{10}$$

where $\mathcal{C}$ is the set of all object classes, $\overline{\Delta_{J_c}}$ is the Lovasz extension to the Jaccard index of class c and $\boldsymbol{m}(c)$ is a vector of errors corresponding to each class. The utility of the Lovasz-Softmax loss function is that it acts as a surrogate to direct optimization of the intersection-over-union score, making it a natural choice for the semantic segmentation task.

4.5. Experiment Results on PASCAL VOC

Table 1 shows our experimental results on Pascal VOC benchmark and Fig.7 illustrates some segmentation samples.

4.6. Experiment Results on MRI

Table 2 shows our experimental results on MRI dataset and Fig.6 shows some segmentation samples.

5. Conclusion

We present our experimental results on two challenging and interesting datasets. Instead of simply implementing the 3 different network designs, we also explore various modifications based on transformation and loss functions. We eventually improve the performance of OCR on the MRI dataset by introducing the Lovász-Softmax Loss to the original network.

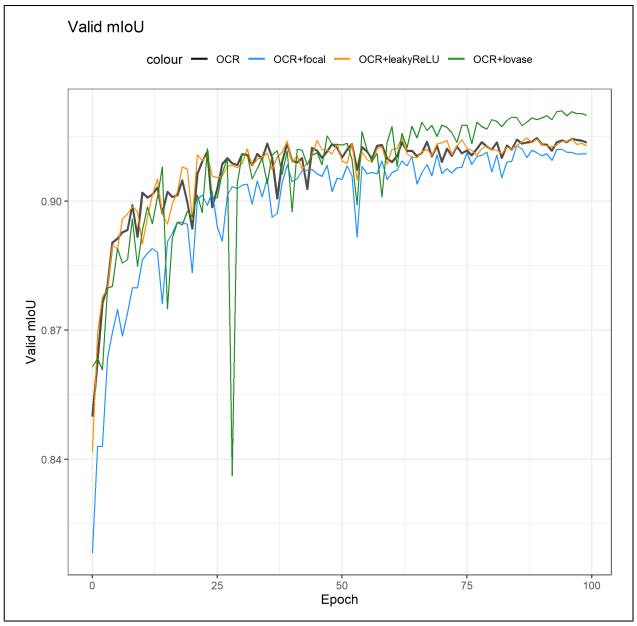

Figure 5. Performance based on MIoU among OCR models.

6. Future Work

Semantic segmentation is a fascinating research topic with numerous applications found in our daily life. We will keep track of the state of the art research related to it. The current frontier works are mainly divided into two types: 1) trying to find a better feature representation, 2) designing a more suitable optimization objective function. We will try to make breakthroughs from these two aspects: 1) Using the idea of self-supervision, we can introduce a huge amount of unlabeled data to improve the feature representation. 2) The current objective function cannot explicitly focus on small targets or very detailed edge information, we may design some loss functions to address this problem.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] M. Berman and M. B. Blaschko. Optimization of the jaccard index for image segmentation with the lovász hinge. *CoRR*, abs/1705.08790, 2017.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[4] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[5] B. Lee, N. Yamanakkanavar, and J. Choi. Automatic segmentation of brain mri using a novel patch-wise u-net deep architecture. *PLoS ONE*, 15, 2020.

[6] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[7] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[9] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang. High-resolution representations for labeling pixels and regions. *CoRR*, abs/1904.04514, 2019.

[10] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015.

[11] Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation. *CoRR*, abs/1909.11065, 2019.

[12] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[13] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han, and E. Ding. Acfnet: Attentional class feature network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6798–6807, 2019.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

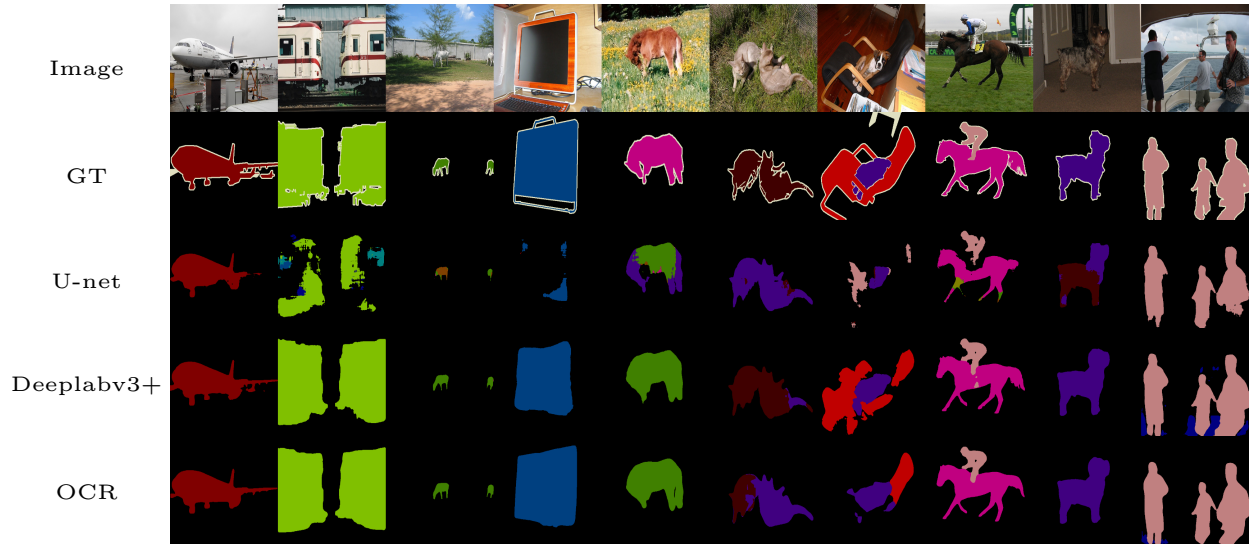

Figure 6. Results on Pascal VOC val samples.

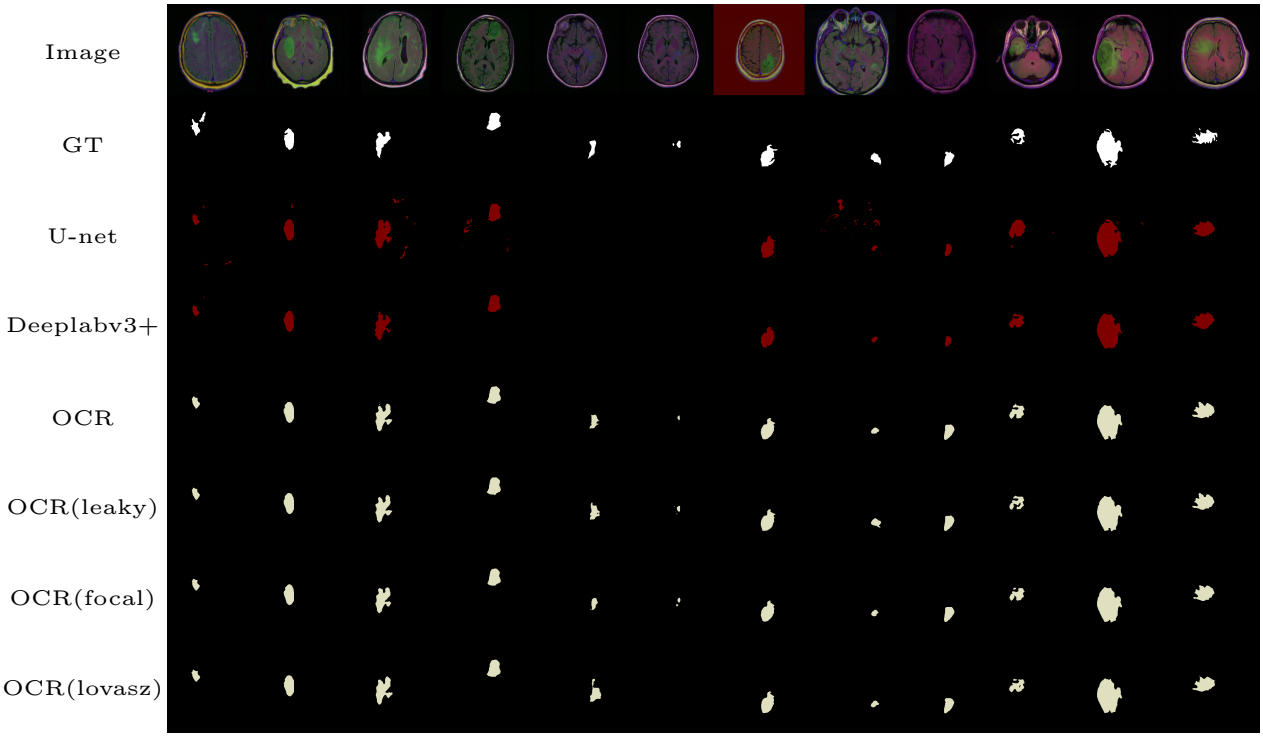

Figure 7. Results on MRI test samples.