End-to-End Boundary Aware Networks for Medical Image Segmentation

Ali Hatamizadeh[1,2], Demetri Terzopoulos[1], and Andriy Myronenko[2]

[1] Computer Science Department, University of California, Los Angeles, CA, USA
[2] NVIDIA, Santa Clara, CA, USA

Abstract. Fully convolutional neural networks (CNNs) have proven to be effective at representing and classifying textural information, thus transforming image intensity into output class masks that achieve semantic image segmentation. In medical image analysis, however, expert manual segmentation often relies on the boundaries of anatomical structures of interest. We propose boundary aware CNNs for medical image segmentation. Our networks are designed to account for organ boundary information, both by providing a special network edge branch and edge-aware loss terms, and they are trainable end-to-end. We validate their effectiveness on the task of brain tumor segmentation using the BraTS 2018 dataset. Our experiments reveal that our approach yields more accurate segmentation results, which makes it promising for more extensive application to medical image segmentation.

Keywords: Medical Image Segmentation · Semantic Segmentation · Convolutional Neural Networks · Deep Learning

1 Introduction

Deep learning approaches to semantic image segmentation have achieved state-of-the-art performance in medical image analysis [9, 8, 5, 4]. With the advent of convolutional neural networks (CNNs), the earliest segmentation methods attempted to classify every pixel based on a corresponding image patch, which often resulted in slow inference times. Fully convolutional neural networks [9], can segment the whole image at once, but the underlying assumption remained—instead of a patch, the corresponding image region (receptive field) centered on the pixel is used for the final pixel segmentation. Since convolutions are spatially invariant, segmentation networks can operate on any image size and infer dense pixel-wise segmentation.

Geirhos et al. [3] empirically demonstrated that, unlike the human visual system, common CNN architectures are biased towards recognizing image textures, not object shape representations. In medical image analysis, however, expert manual segmentation usually relies on boundary and organ shape identification. For instance, a radiologist segmenting a liver from CT images would usually trace liver edges first, from which the internal segmentation mask is easily deduced.

This observation motivates us to devise segmentation networks that prioritize the representation of edge information in anatomical structures by leveraging an additional edge module whose training is supervised by edge-aware loss functions.

Recently, several authors have pursued deep learning approaches for object edge prediction. Yu et al. [11] proposed a multilabel semantic boundary detection network to improve a wide variety of vision tasks by predicting edges directly, including a new skip-layer architecture in which category-wise edge activations at the top convolution layer share and are fused with the same set of bottom layer features, along with a multilabel loss function to supervise the fused activations. Subsequently, Yu et al. [12] showed that label misalignment can cause considerably degraded edge learning quality, and addressed this issue by proposing a simultaneous edge alignment and learning framework. Acuna et al. [1] predicted object edges by identifying pixels that belong to class boundaries, proposing a new layer and a loss that enforces the detector to predict a maximum response along the normal direction at an edge, while also regularizing its direction. Takikawa et al. [10] proposed gated-shape CNNs for semantic segmentation of natural images in which such gates are employed to remove the noise from higher-level activations and process the relevant boundary-related information separately. Aiming to learn semantic boundaries, Hu et al. [6] presented a framework that aggregates different tasks of object detection, semantic segmentation, and instance edge detection into a single holistic network with multiple branches, demonstrating significant improvements over conventional approaches through end-to-end training.

In the present paper, we introduce an encoder-decoder architecture that leverages a special interconnected edge layer module that is supervised by edge-aware losses in order to preserve boundary information and emphasize it during training. By explicitly accounting for the edges, we encourage the network to internalize edge importance during training. Our method utilizes edge information only to assist training for semantic segmentation, not for the main purpose of predicting edges directly. This strategy enables a structured regularization mechanism for our network during training and results in more accurate and robust segmentation performance during inference. We validate the effectiveness of our network on the task of brain tumor segmentation using the BraTS 2018 dataset [2].

2 Methods

2.1 Architecture

Our network comprises a main encoder-decoder stream for semantic segmentation as well as a shape stream that processes the feature maps at the boundary level (Fig. 1). In the encoder portion of the main stream, every resolution level includes two residual blocks whose outputs are fed to the corresponding resolution of the shape stream. A 1×1 convolution is applied to each input to the shape stream and the result is fed into an attention layer that is discussed in the next section. The outputs of the first two attention layers are fed into connection residual blocks. The output of the last attention layer is concatenated with the output of

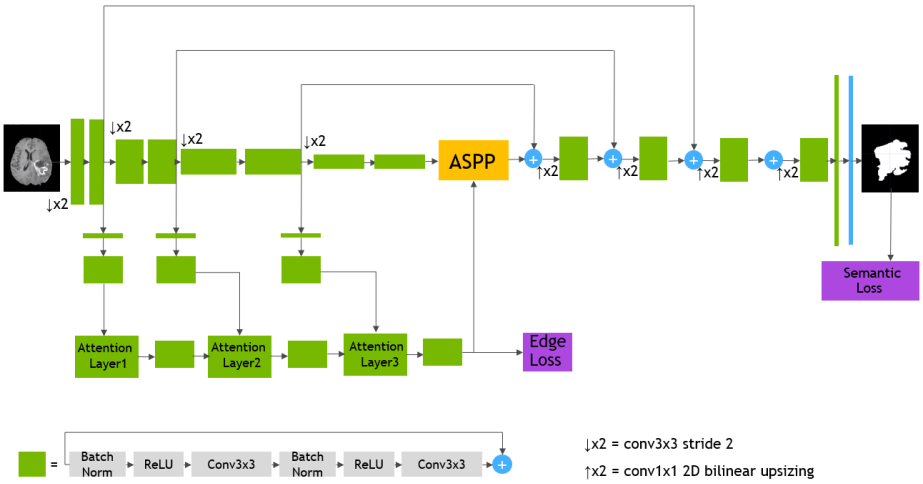


Fig. 1: Our 2D fully convolutional architecture. We use dilated spatial pyramid pooling to effectively aggregate the outputs of different stages.

the encoder in the main stream and fed into a dilated spatial pyramid pooling layer. Losses that contribute to tuning the weights of the model come from the output of the shape stream that is resized to the original image size, as well as the output of the main stream.

2.2 Attention Layer

Each attention layer receives inputs from the previous attention layer as well as the main stream at the corresponding resolution. Let s_l and m_l denote the attention layer and main stream layer inputs at resolution l. We first concatenate s_l and m_l and apply a 1×1 convolution layer $C_{1\times1}$ followed by a sigmoid function σ to obtain an attention map

$$\alpha_l = \sigma\big(C_{1\times1}(s_l \parallel m_l)\big). \tag{1}$$

An element-wise multiplication is then performed with the input to the attention layer to obtain the output of the attention layer, denoted as

$$o_l = s_l \odot \alpha_l. \tag{2}$$

2.3 Boundary Aware Segmentation

Our network jointly learns the semantics and boundaries by supervising the output of the main stream as well as the edge stream. We use the generalized Dice loss on predicted outputs of the main stream and the shape stream. Additionally, we add a weighted binary cross entropy loss to the shape stream loss in order to

deal with the large imbalance between the boundary and non-boundary pixels. The overall loss function of our network is

$$L_{\text{total}} = \lambda_1 L_{\text{Dice}}(y_{\text{pred}}, y_{\text{true}}) + \lambda_2 L_{\text{Dice}}(s_{\text{pred}}, s_{\text{true}}) + \lambda_3 L_{\text{Edge}}(s_{\text{pred}}, s_{\text{true}}), \quad (3)$$

where y_{pred} and y_{true} denote the pixel-wise semantic predictions of the main stream while s_{pred} and s_{true} denote the boundary predictions of the shape stream; s_{true} can be obtained by computing the spatial gradient of y_{true}.

The Dice loss [7] in (3) is

$$L_{\text{Dice}} = 1 - \frac{2\sum y_{\text{true}} y_{\text{pred}}}{\sum y_{\text{true}}^2 + \sum y_{\text{pred}}^2 + \epsilon}, \quad (4)$$

where summation is carried over the total number of pixels and ϵ is a small constant to prevent division by zero.

The edge loss in (3) is

$$L_{\text{Edge}} = -\beta \sum_{j \in y_+} \log P(y_{\text{pred},j} = 1|x;\theta) - (1-\beta) \sum_{j \in y_-} \log P(y_{\text{pred},j} = 0|x;\theta), \quad (5)$$

where x, θ, y_-, and y_+ denote the input image, CNN parameters, and edge and non-edge pixel sets, respectively, β is the ratio of non-edge pixels over the entire number of pixels, and $P(y_{\text{pred},j})$ denotes the probability of the predicated class at pixel j.

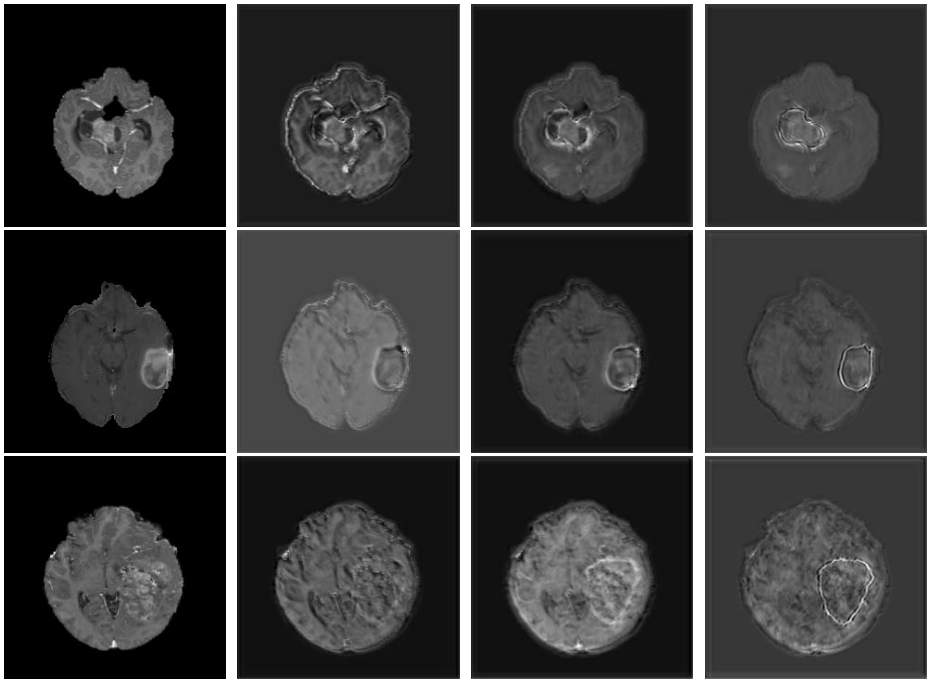
3 Experiments

3.1 Datasets

In our experiments, we used the BraTS 2018 [2], which provides multimodal 3D brain MRIs and ground truth brain tumor segmentations annotated by physicians, consisting of 4 MRI modalities per case (T1, T1c, T2, and FLAIR). Annotations include 3 tumor subregions—the enhancing tumor, the peritumoral edema, and the necrotic and non-enhancing tumor core. The annotations were combined into 3 nested subregions—whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The data were collected from 19 institutions, using various MRI scanners. For simplicity, we use only a single input MRI modality (T1c) and aim to segment a single tumor region—TC, which includes the main tumor components (nectrotic core, enhancing, and non-enhancing tumor regions). Furthermore, even though the original data is 3D ($240 \times 240 \times 155$), we operate on 2D slices for simplicity. We have extracted several axial slices centered around the tumor region from each 3D volume, and combined them into a new 2D dataset.

3.2 Implementation Details

We have implemented our model in Tensorflow. The brain input images were resized to predefined sizes of 240×240 and normalized to the intensity range



(a) (b) (c) (d)

Fig. 2: (a) Input image. Outputs of : (b) Attention Layer 1. (c) Attention Layer 2. (d) Attention Layer 3.The boundary emphasis becomes more prominent in the subsequent attention layers.

$[0, 1]$. The model was trained on NVIDIA Titan RTX and an Intel Core i7-7800X CPU @ 3.50GHz $\times$ 12 with a batch size of 30 for all models. We used $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.1$ in (3). The Adam optimization algorithm was used with initial learning rate of $\alpha_0 = 1.0^{-3}$ and further decreased according to

$$\alpha = \alpha_0 \left(1 - e/N_e\right)^{0.9}, \tag{6}$$

where e denotes the current epoch and N_e the total number of epochs, following [8]. We have evaluated the performance of our model by using the Dice score, Jaccard index, and Hausdorff distance.

4 Results and Discussion

Boundary Stream: Fig. 2 demonstrates the output of each of the attention layers in our dedicated boundary stream. In essence, each attention layer progressively localizes the tumor and refines the boundaries. The first attention layer has learned rough estimate of the boundaries around the tumor and localized it, whereas the second and third layers have learned more fine-grained details of the

Model	Dice Score	Jaccard Index	Hausdorff Distance
U-Net [9]	0.731±0.230	0.805 ±0.130	3.861±1.342
V-Net [7]	0.769±0.270	0.837±0.140	3.667±1.329
Ours (no edge loss)	0.768±0.236	0.832±0.136	3.443±1.218
Ours	**0.822±0.176**	**0.861±0.112**	**3.406±1.196**

Table 1: Performance evaluations of different models. We validate the contribution of the edge loss by measuring performance with and without this layer

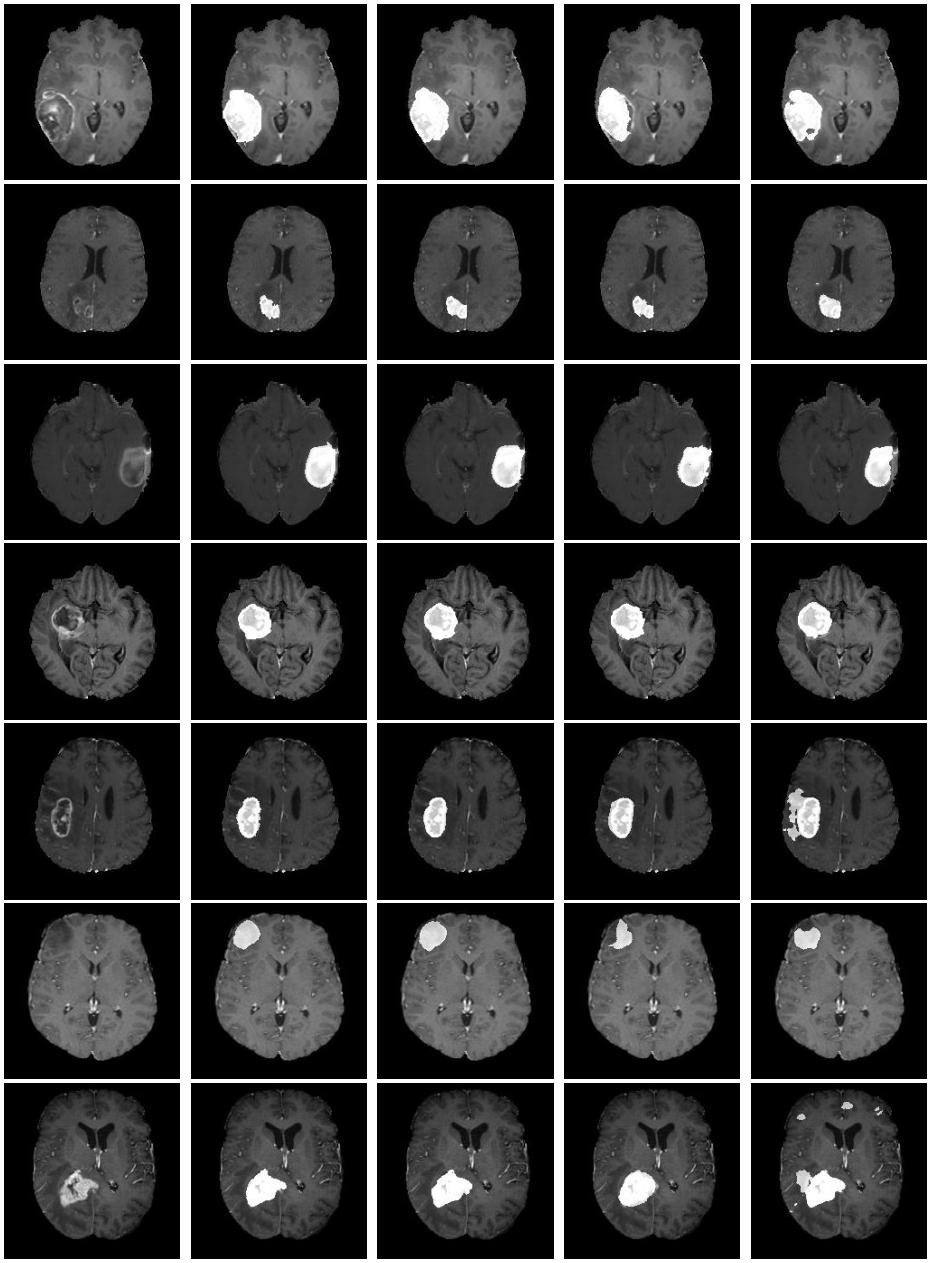
edges and boundaries, refining the localization. Moreover, since our architecture leverages a dilated spatial pyramid pooling to merge the learned feature maps of the regular segmentation stream and the boundary stream, multiscale regional and boundary information have been preserved and fused properly, which has enabled our network to capture the small structural details of the tumor.

Edge-Aware Losses: To validate the effectiveness of the loss supervision, we have trained our network without enforcing the supervision of the edge loss during the learning process, but with the same architecture. Table 1 shows that our network performs very similarly to V-Net [7] without edge supervision, since ours employs similar residual blocks as V-Net in its main encoder-decoder, and its boundary stream does not seem to contribute to the learning of useful features for segmentation. In essence, the boundary stream also impacts the down-stream layers of the encoder by emphasizing edges during training.

Comparison to Competing Methods: We have compared the performance of our model against the most popular deep learning-based semantic segmentation networks, U-Net [9] and V-Net [7] (Fig. 3). Our model outperforms both by a considerable margin in all evaluation metrics. In particular, U-Net performs poorly in most cases due to the high false positive of its segmentation predictions, as well as the imprecision of its boundaries. The powerful residual block in the V-Net architecture seems to alleviate these issues to some extent, but V-Net also fails to produce high-quality boundary predictions. The emphasis of learning useful edge-related information during the training of our network appears to effectively regularize the network such that boundary accuracy is improved.

5 Conclusion

We have proposed an end-to-end-trainable boundary aware network for joint semantic segmentation of medical images. Our network explicitly accounts for object edge information by using a dedicated shape stream that processes the feature maps at the boundary level and fuses the multiscale contextual information of the boundaries with the encoder output of the regular segmentation stream. Additionally, edge-aware loss functions emphasize learning of the edge information



(a) (b) (c) (d) (e)

Fig. 3: (a) Input images. (b) Labels. (c) Ours. (d) V-Net. (e) U-Net.

during training by tuning the weights of the downstream encoder and regularizing the network to prioritize boundaries. We have validated the effectiveness of our approach on the task of brain tumor segmentation using the BraTS 2018 dataset. Our results indicate that our network produces more accurate segmentation outputs with fine-grained boundaries in comparison to the popular segmentation networks U-Net and V-Net.

References

[1] Acuna, D., Kar, A., Fidler, S.: Devil is in the edges: Learning semantic boundaries from noisy annotations. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[2] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data 4 (2017)

[3] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (ICLR) (2019)

[4] Hatamizadeh, A., Hoogi, A., Sengupta, D., Lu, W., Wilcox, B., Rubin, D., Terzopoulos, D.: Deep active lesion segmentation. arXiv preprint arXiv:1908.06933 (2019)

[5] Hatamizadeh, A., Hosseini, H., Liu, Z., Schwartz, S.D., Terzopoulos, D.: Deep dilated convolutional nets for the automatic segmentation of retinal vessels. arXiv preprint arXiv:1905.12120 (2019)

[6] Hu, Y., Zou, Y., Feng, J.: Panoptic edge detection. https://arxiv.org/abs/1906.00590 (2019)

[7] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision (3DV) (2016)

[8] Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: BrainLes, Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 311–320. LNCS, Springer (2018)

[9] Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI. LNCS, vol. 9351, pp. 234–241 (2015)

[10] Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. arXiv preprint arXiv:1907.05740 (2019)

[11] Yu, Z., Feng, C., Liu, M., Ramalingam, S.: Casenet: Deep category-aware semantic edge detection. In: CVPR (2017)

[12] Yu, Z., Liu, W., Zou, Y., Feng, C., Ramalingam, S., Vijaya Kumar, B., Kautz, J.: Simultaneous edge alignment and learning. In: European Conference on Computer Vision (ECCV) (2018)