Chapter 6

Applications of the ESPNet architecture in medical imaging

Sachin Mehta[1]

Nicholas Nuechterlein[1]

Ezgi Mercan[2]

Beibin Li[1]

Shima Nofallah[1]

Wenjun Wu[1]

Ximing Lu[1]

Anat Caspi[1]

Mohammad Rastegari (Change this to 1,)[3]

Joann Elmore[4]

Hannaneh Hajishirzi[1]

Linda Shapiro[1]

[1]University of Washington, Seattle, WA, United States

[2]Seattle Children's Hospital, Seattle, WA, United States

[3]Allen Institute of Artificial Intelligence (AI2) and XNOR.AI

[4]University of California, Los Angeles, CA, United States

6.1 Introduction

Medical imaging creates visual representations of the human body, including organs and tissues, to aid in diagnosis and treatment. These visual representations are effective in a variety of medical settings and have become integral in clinical decision making. A common example is X-ray-based radiography examinations that are used to capture visual representations of the interior structure of the body. Other examples include PAP smear analysis for cervical cancer screening and whole slide biopsy analysis for multiple kinds of cancer, including breast cancer and melanoma. Today's human physician is no longer able to interpret the vast amount of information now available in medical imaging. For example, there are hundreds of thousands of cells on some of the whole slide images (WSIs) of a biopsy.

Machine learning is an effective tool that enables machines (or computers) to learn meaningful patterns from medical imaging data, which can be used to build computer-aided diagnostic systems [1,2]. With the recent advancements in hardware technology and the availability of a large amount of medical imaging data, deep learning-based methods, especially convolutional neural networks (CNNs) [3], are gaining attention in medical image analysis [4,5]. Researchers have applied deep learning-based methods to a variety of medical data, such as WSIs [6], magnetic resonance (MR) tumor scans [7], electron microscopic recordings [8], and tasks, such as cancer diagnosis [2,9] and cellular-level entity detection [8,10].

In this chapter, we will describe the ESPNet architecture [11] that has been successfully applied across a variety of visual recognition tasks, including image classification, object detection, semantic segmentation, and medical image analysis [6,12,13]. To demonstrate the modeling power of the ESPNet architecture, we study the application of ESPNet to two different medical imaging tasks: (1) tissue-level segmentation of breast biopsy WSIs and (2) tumor segmentation in 3D brain MR images. Our results show that the ESPNet architecture learns meaningful representations efficiently that allows it to deliver good

accuracy on different tasks.

The rest of the chapter is organized as follows. Section 6.2 reviews the different types of convolutions that are used in the ESPNet architecture, which is described in Section 6.3. Experimental results on two different medical imaging datasets are provided in Section 6.4. Section 6.5 concludes the chapter.

6.2 Background

The convolution operation lies at the heart of many computer vision algorithms, including CNNs. In this section, we will briefly review two types of convolutions, standard and dilated convolutions, which are used in the ESPNet architecture.

6.2.1 Standard convolution

For a given 2D input image $\boldsymbol{X}$, the standard convolution moves a kernel $\boldsymbol{K}$ of size $n \times n$[1] over every spatial location of the input $X(i,j)$ to produce the output $\boldsymbol{Y}$. Mathematically, it can be defined as:

$$\boldsymbol{Y}(i,j) = \boldsymbol{X} * \boldsymbol{K} = \sum_{u=-\left\lfloor \frac{n}{2} \right\rfloor}^{\left\lfloor \frac{n}{2} \right\rfloor} \sum_{v=-\left\lfloor \frac{n}{2} \right\rfloor}^{\left\lfloor \frac{n}{2} \right\rfloor} \boldsymbol{X}(i,j)\boldsymbol{K}(i+u, j+v),$$

where * denotes the convolution operations.

6.2.2 Dilated convolution

Dilated convolutions[2] are a special form of standard convolutions in which holes are inserted between kernel elements to increase the effective receptive field. These convolutions have been widely used in semantic segmentation networks [14,15]. For a given 2D input image $\boldsymbol{X}$, the dilated convolution moves a kernel $\boldsymbol{K}$ of size $n \times n$ with a dilation rate of r over every spatial location of the input $X(i,j)$ to produce the output $\boldsymbol{Y}$. Mathematically, it can be defined as:

$$Y(i,j) = X *_r K = \sum_{u=-\left\lfloor \frac{n}{2} \right\rfloor}^{\left\lfloor \frac{n}{2} \right\rfloor} \sum_{v=-\left\lfloor \frac{n}{2} \right\rfloor}^{\left\lfloor \frac{n}{2} \right\rfloor} X(i,j)K(i+ur, j+vr),$$

where $*_r$ denotes the dilated convolution operation with a dilation rate of r. The effective receptive field of a dilated convolutional kernel $\boldsymbol{K}$ of size $n \times n$ with a dilation rate of r is $((n-1)r+1) \times ((n-1)r+1)$. Note that dilated convolutions are the same as standard convolutions when the dilation rate r is 1. An example comparing dilated and standard convolutions is shown in Fig. 6.1.

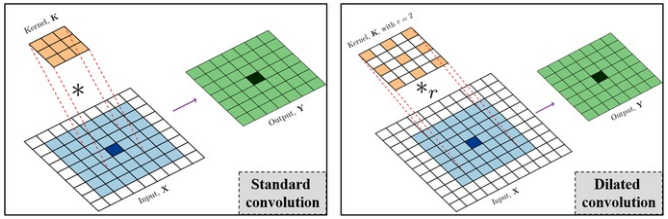

Figure 6.1 A comparison between standard and dilated convolution operations. Input $\boldsymbol{X}$ is padded with zeros (white cells) so that the resulting output after the convolution operation is of the same size as that of the input.

6.3 The ESPNet architecture

In this section, we elaborate on the details of the ESPNet architecture. We first describe the ESP module, the core building block of the ESPNet architecture, and then describe different variants of the ESPNet architecture for analyzing different types of medical images, such as whole slide biopsy images and 3D brain tumor images.

6.3.1 Efficient spatial pyramid unit

The core building block of the ESPNet architecture is the efficient spatial pyramid (ESP) unit. To be efficient, the ESP unit decomposes the standard convolution into a point-wise convolution and spatial pyramid of dilated

convolutions using the RSTM (reduce, split and transform, and merge) principle:

- Reduce: The ESP unit applies d point-wise[3] (or 1×1) convolutions to an input tensor $\boldsymbol{X} \in \mathbb{R}^{N \times H \times W}$ to produce an output tensor $\boldsymbol{X_P} \in \mathbb{R}^{d \times H \times W}$, where W and H represent the width and height of the tensor, whereas N and d represent the number of channels in the input and output tensor.
- Split and Transform: To learn the spatial representations from a large receptive field efficiently, the ESP unit applies K, $n \times n$ dilated convolutions *simultaneously* with different dilation rates $r_k = 2^k$ to $\boldsymbol{X_P} \in \mathbb{R}^{d \times H \times W}$ to produce output $\boldsymbol{X_D^k} \in \mathbb{R}^{d \times H \times W}$, $k = \{0, \ldots, K-1\}$. To learn spatial representations from a large receptive field, the ESP unit uses a different dilation rate, $r_k = 2^k$, in each branch. This allows the ESP unit to learn spatial representations from an effective receptive field of $(n-1)2^{K-1}+1$, where $n \times n$ denotes the kernel size.
- Merge: The ESP unit concatenates these d-dimensional feature maps to produce a M-dimensional feature map $\boldsymbol{Y} = \{\boldsymbol{X_D^0} \| \boldsymbol{X_D^1} \| \ldots \| \boldsymbol{X_D^{K-1}}\}$, where $\|$ represents the concatenation operation and $M = Kd$.

Fig. 6.2 compares the ESP unit with the standard convolution. The ESP unit learns $\mathrm{d}N + n^2\mathrm{d}^2K = \frac{M(N+n^2M)}{K}$ parameters and performs $\frac{WHM(N+n^2M)}{K}$ operations. Compared to n^2NM parameters and n^2NMWH operations for the standard convolution, the RSTM principle reduces the total number of parameters and operations by a factor of $\frac{n^2NK}{N+n^2M}$, while simultaneously increasing the effective receptive field by approximately $2^{K-1} \times 2^{K-1}$.

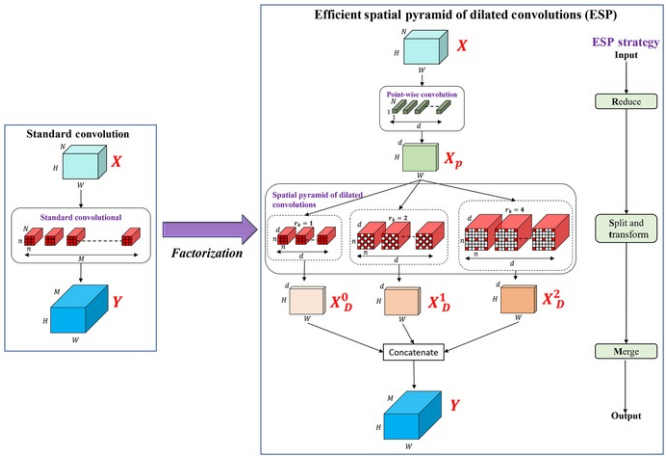

Figure 6.2 A comparison between the standard convolution and the ESP unit. In the ESP unit, the standard convolution is decomposed (We only refer to white color in this figure caption, which should be okay for both color and B&W prints. Same for Figure 6.1) into a point-wise convolution and a spatial pyramid of dilated convolutions using the RSTM principle. Note that in dilated convolutions, zeros (represented in white) are inserted between the kernel elements to increase the receptive field.

6.3.1.1 Hierarchical feature fusion for degridding in the efficient spatial pyramid unit

Dilated convolutions insert zeros, controlled by the dilation rate, between kernel elements to increase the receptive field of the kernel. However, this introduces unwanted checkerboard or gridding artifacts on the output, as shown in Fig. 6.3. The ESP unit introduces a hierarchical feature fusion (HFF) method to remove these artifacts in a computationally efficient manner[4]. HFF hierarchically adds outputs I_D^k and then concatenate them to produce a high-dimensional feature map $\hat{\boldsymbol{Y}} = \{\boldsymbol{X_D^0} \| \boldsymbol{X_D^0} \oplus \boldsymbol{X_D^1} \| \ldots \| \boldsymbol{X_D^0} \oplus \boldsymbol{X_D^1} \oplus \cdots \oplus \boldsymbol{X_D^{K-1}}\}$, where $\oplus$ represents the element-wise addition operation. Fig. 6.4 visualizes the ESP unit with and without HFF.

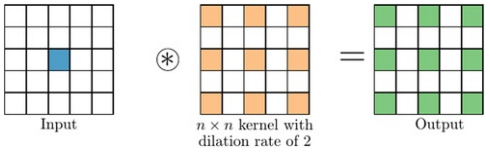

Figure 6.3 This figure illustrates a gridding artifact example, where a 5×5 input with single-active pixel (represented in blue) is convolved with a 3×3 dilated convolution kernel to produce an output with a gridding artifact. Active kernel elements are represented in ora (This figure illustrates a gridding artifact example, where a 5x5 input with a single-active pixel is convolved with a 3x3 dilated convolution kernel to produce an ouput with a gridding artifact. Active input, output, and kernel elements are shown in color.) nge.

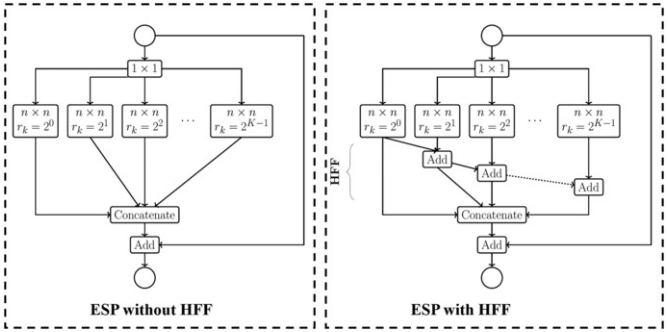

Figure 6.4 Block level visualization of the ESP unit without and with hierarchical feature fusion (HFF).

6.3.2 Segmentation architecture

Most image segmentation architectures follow an encoder–decoder structure [16]. The encoder network is a stack of encoding units, such as the bottleneck unit in ResNet [17], and downsampling units that help the network learn multiscale representations. Spatial information is lost during filtering and downsampling operations in the encoder; the decoder tries to recover the loss of this information. The decoder can be viewed as an inverse of the encoder that stacks upsampling and decoding units to learn representations that can help produce either binary or multiclass segmentation masks. It is important to note that a vanilla encoder–decoder network does not share information between encoding and decoding units at each spatial level. To share information between encoding and decoding units at each spatial level, U-Net [8] introduces a skip connection between the encoding and decoding units at each spatial level. This connection establishes a direct link between the encoding and the decoding units and improves gradient flow, thus improving segmentation performance. Fig. 6.5 compares the vanilla encoder–decoder and U-Net style encoder–decoder architectures.

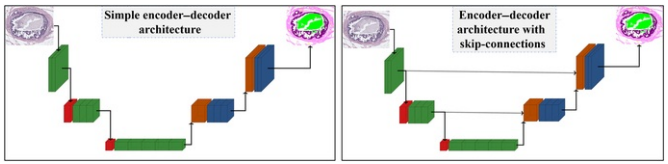

Figure 6.5 Encoder–decoder architectures for segmentation. Colors represent different types of units in the encoder–decoder arachitecture (Encoder-decoder architectures for tissue-level semantic segmentation): encoding unit in green, downsampling unit in red, upsampling unit in orange, and decoding unit in blue.

The ESPNet architecture for semantic segmentation extends the U-Net. In contrast to using computationally expensive VGG-style [18] blocks for encoding and decoding the information, the ESPNet architecture uses computationally efficient ESP units.

6.4 Experimental results

In this section, we provide results of the ESPNet architecture on two different medical imaging datasets: (1) breast biopsy whole slide imaging and (2) brain tumor segmentation.

6.4.1 Breast biopsy whole slide image dataset

6.4.1.1 Dataset

The breast biopsy dataset consists of 240 WSIs with haematoxylin and eosin (H&E) staining [19,20]. Three expert pathologists independently interpreted each of the 240 cases and then met to review each case together and provide a consensus reference diagnosis label, which we use as the gold standard ground truth for each case. Additionally, the expert pathologists marked 428 region of interests (ROIs) on these 240 WSIs that were representative of their diagnosis. Out of these 428 ROIs, 58 were manually segmented by an expert into eight different tissue-level segmentation labels: background, benign epithelium, malignant epithelium, normal stroma, desmoplastic stroma, secretion, blood, and necrosis. Fig. 6.6 visualizes some ROIs along with their tissue-level segmentation labels. Furthermore, we split these 58 ROIs into two equally sized subsets, a training set (30 ROIs) and a test set (28 ROIs), while keeping the distribution of diagnostic categories similar (see Table 6.1). A total of 87 pathologists who were actively interpreting breast biopsies in their own clinical practices participated in the study and provided one of the four

diagnostic labels (benign, atypia, ductal carcinoma in situ, and invasive cancer) per WSI of a case. Each pathologist provided diagnostic labels for a randomly assigned subset of 60 patients' WSIs, producing an average of 22 diagnostic labels per case.

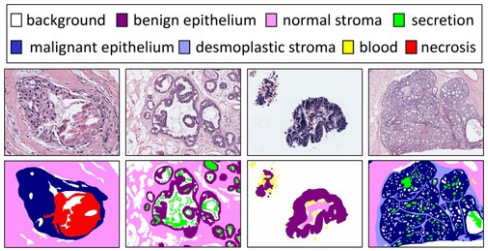

Figure 6.6 The set of ROIs (top row) along with their tissue-level segmentation labels (bottom row) from the dataset. (We can't change colors for tissue labels in Figure 6.6., 6.7, 6.8, and 6.9)

Table 6.1 Diagnostic category-wise distribution of 58 region of interests (ROIs) for the segmentation subset.

	# ROI			Average ROI size (in pixels)
Expert consensus reference diagnostic category	Training	Test	Total	
Benign	4	5	9	$9K \times 9K$
Atypia	11	11	22	$6K \times 7K$
DCIS	12	10	22	$8K \times 10K$
Invasive	3	2	5	$38K \times 44K$
Total	30	28	58	$10K \times 12K$

6.4.1.2 Training

The breast biopsy ROIs (or WSIs) have huge spatial dimensions, often of the order of gigapixels. Due to such high spatial dimensionality of these images and the limited computational capability of existing hardware resources, it is difficult to train conventional CNNs directly on these images. Therefore we follow a patch-based approach that splits a ROI (or WSI) into small patches of size 384×384 with an overlap of 56 pixels between consecutive patches. We use standard augmentation strategies such as random flipping, cropping, and resizing to prevent overfitting. For training, we split the training set of 30 ROIs into training and validation subsets with a 90:10 ratio. We train our network for 100 epochs using stochastic gradient descent (SGD) with an initial learning rate of 0.0001, which is decreased by 0.5 after every 30 epochs. To evaluate the tissue-level segmentation performance, we use mean intersection over union (mIOU), a widely used metric to evaluate segmentation performance.

The shape and structure of objects of interest in the breast biopsy are variable in size, and splitting such structures into fixed size using a patch-based approach limits the contextual information; CNNs tend to make segmentation errors especially at the border of the patch (see Fig. 6.7). To avoid such errors, we centrally crop a 384×384 prediction to a size of 256×256, as shown in Fig. 6.8.

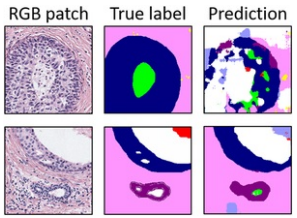

Figure 6.7 Segmentation errors along the border of the patch.

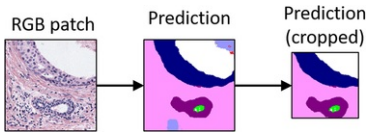

Figure 6.8 Center cropping to minimize segmentation errors along patch border.

6.4.1.3 Segmentation results

Tables 6.2 and 6.3 summarize the performance of different segmentation architectures.

Table 6.2 Impact of skip connections and different decoding units.

Encoding–decoding units	Skip connection		Network parameters	mIOU
	Add	Concat		
ESP–ESP (Could you please edit the layout of Table 6.2, so that it is the same as we submited? In particular, merge the three rows in Column 1 (corresponding to ESP-ESP setting), so that readers may not get confused with which row is for which setting)			1.95 M	35.23
	✓		1.95 M	36.19
		✓	2.25 M	38.03
ESP–PSP		✓	2.75 M	44.03

Table 6.3 Comparison with state-of-the-art methods.

Method	Network parameters	mIOU
Superpixel + SVM	NA	25.8
SegNet	12.80 M	37.6
SegNet + additive skip connections	12.80 M	38.1
Multiresolution	26.03 M	44.20
Y-Net (ESP–PSP)	2.75 M	44.03

Notes: The results in the first four rows are taken from Refs [6,11,21].

6.4.1.4 Skip connections

Skip connections between the encoder and the decoder in Fig. 6.5 can be constructed using two operations: (1) element-wise addition and (2) concatenation. The impact of these connections is studied in Table 6.2. We can see that encoder–decoder networks with skip connection constructed using concatenation operations improve the accuracy of the vanilla encoder–decoder by about 4% and that the encoder–decoder networks with skip connections constructed using element-wise addition operations improve the accuracy of the vanilla encoder–decoder by about 2%.

6.4.1.5 Pyramidal spatial pooling as a decoding unit

Traditionally, the decoding unit in encoder–decoder architectures, such as SegNet [16] and U-Net [8], is the same as the encoding unit. In our recent work [6,11,21], we introduced a general encoder–decoder network, sketched in Fig. 6.5, that allows the use of different encoding and decoding units. When we replaced the ESP unit with the pyramidal spatial pooling (PSP) unit [22] in the decoder, the performance of our network improved by about 6%. This is likely because the pooling operations in the PSP unit allow the capture of better global contextual representations (Fig. 6.9).

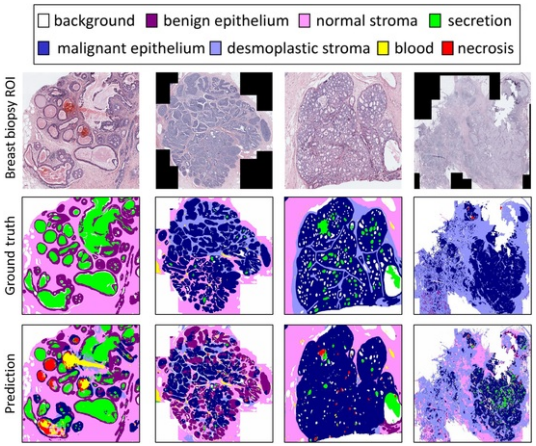

Figure 6.9 ROI-wise predictions. The first row shows different breast biopsy ROIs, the second row shows tissue-level ground-truth labels; the third row shows predictions. *ROI*, region of interest.

6.4.1.6 Comparison with state-of-the-art methods

Table 6.3 compares the performance with different methods. We can see that the asymmetric encoder–decoder structure, with the ESP as the encoding unit and the PSP as the decoding unit, allows us to build a light-weight network that has $9.5\times$ fewer parameters than the network of Refs [6,11,21] while delivering similar segmentation performance.

6.4.1.7 Tissue-level segmentation masks for computer-aided diagnosis

Tissue-level segmentation masks provide a powerful abstraction for diagnostic classification. To demonstrate the descriptive power of tissue-level segmentation masks, we extract and study the impact of two features, tissue-distribution and structural features [2], for a four-class breast cancer diagnostic task. For this analysis, we use the 428 ROIs. We use a support-vector machine classifier with a polynomial kernel in a leave-one-out-cross-validation strategy. We measure the performance in terms of sensitivity, specificity, and accuracy. Table 6.4 summarizes diagnostic classification results. We can see that simple features extracted from tissue-level segmentation masks are powerful, and we are able to attain accuracies similar to pathologists.

Table 6.4 Impact of different features on diagnostic classification from Ref. [2].

Diagnostic features	Accuracy	Sensitivity	Specificity
Invasive versus noninvasive			
Tissue-distribution feature	0.94	0.70	0.95
Structural feature	0.91	0.49	0.96
Pathologists	0.98	0.84	0.99
Atypia and DCIS versus benign			
Tissue-distribution feature	0.70	0.79	0.41
Structural feature	0.70	0.85	0.45
Pathologists	0.81	0.72	0.62
DCIS versus atypia			
Tissue-distribution feature	0.83	0.88	0.78
Structural feature	0.85	0.89	0.80

Pathologists	0.80	0.70	0.82

Notes: The best numbers are indicated in bold.

6.4.2 Brain tumor segmentation

6.4.2.1 Dataset

The Multimodal Brain Tumor Segmentation Challenge (BraTS) 2018 training set consists of 285 multi(multi-instituitional?)institutional preoperative multimodal MR tumor scans, each consisting of T1, postcontrast T1-weighted (T1ce), T2, and FLAIR volumes [23]. Each volume is annotated with voxel labels corresponding to the following tumor compartments: enhancing tumor, peritumoral edema, background, and necrotic core and nonenhancing tumor. Necrotic core and nonenhancing tumor share a single label. These data are coregistered to the standard MNI anatomical template, interpolated to the same resolution, and skull-stripped. Ground-truth segmentations are manually drawn by radiologists. Fig. 6.10 visualizes some MR modalities along with their voxel-level segmentation labels.

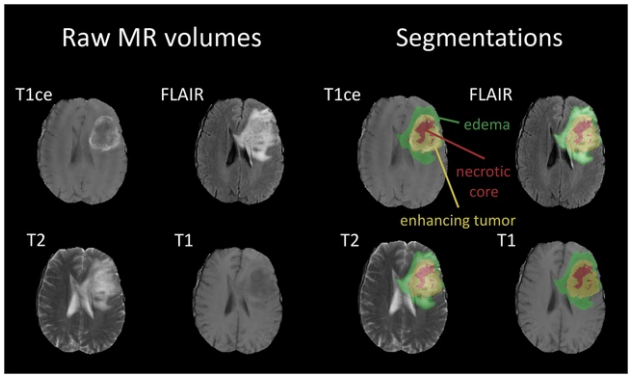

Figure 6.10 The set of MR modalities (left) along with their voxel-level segmentation labels (right). *MR*, magnetic resonance.

6.4.2.2 Training

MR scans are high-dimensional: each of the four MR sequence volumes has a dimension of $240 \times 240 \times 155$. We use standard augmentation strategies such as random flipping, cropping, and resizing to prevent overfitting. For training, we split the training set of 285 MR scans into train and validation subsets with 80:20 ratio (228:57). We train our network for 300 epochs using SGD with an initial learning rate of 0.0001 and decreased it to 0.00001 after 200 epochs. We evaluate the segmentation performance using an online server, that measures the segmentation performance in terms of the Dice score. Furthermore, we adapt the ESP module for volume-wise segmentation by replacing the spatial dilated convolutions in Fig. 6.4 with the volume-wise dilated convolutions. For more details, please see Ref. [13].

6.4.2.3 Results

Table 6.5 summarizes the results of the BRATS 2018 online test and validation sets. Unlike the winning entries from the competition that often use specific normalization techniques, model ensembling, and postprocessing methods such as conditional random field (CRF) to boost the performance [7,24], our network was able to attain good segmentation performance while learning merely 3.8 million parameters, which are an order of magnitude fewer than existing methods. Furthermore, our visual inspection reveals convincing performance; we display segmentation results overlaid on different modalities in Fig. 6.11. It is important to note that the predictions made by our network are smooth and lack some of the granularity present in the ground-truth segmentation. This is likely because our model is not able to learn such granular details with a limited amount of training data. We believe that postprocessing techniques, such as CRFs, would help our network in capturing such granular details and we will study such methods in the future.

Table 6.5 Results obtained by our method [13] of the BraTS 2018 online test and validation set.

Networks	Whole tumor	Enhancing tumor	Tumor core
Ours—validation	0.883	0.737	0.814
Ours—test	0.850	0.665	0.782

Myronenko [24]	0.884	0.766	0.815

Notes: We also compare with the winning entry on the test set [24], which uses an ensemble of 10 models and special normalization techniques.

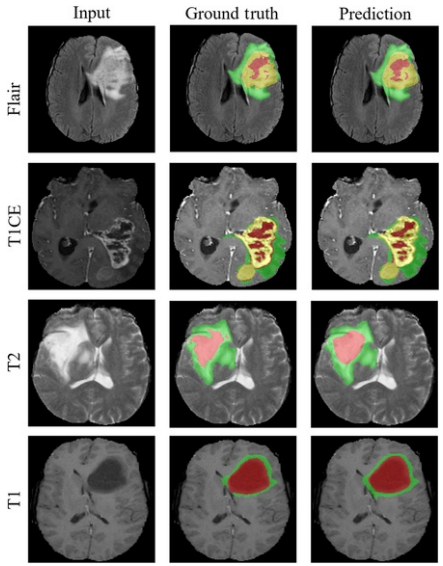

Figure 6.11 This figure visualizes the segmentation results produced by our network overlaid on each of the four MR modalities in the BRATS dataset. *MR*, magnetic resonance.

6.4.3 Other applications

The ESPNet architecture is a general-purpose architecture and can be used across a wide variety of applications, ranging from image classification to language modeling. In our work, we have used the ESPNets for image classification [12], object detection [12], semantic segmentation [6,11,21], language modeling [12], and autism spectral disorder prediction [25].

6.5 Conclusion

This chapter describes the ESPNet architecture that is built using a novel convolutional unit, the ESP unit, introduced in Refs [6,11,21]. To effectively demonstrate the modeling power of the ESPNet architecture on medical imaging, we study it on two different datasets: (1) a breast biopsy WSI dataset and (2) a brain tumor segmentation MR imaging dataset. Our analysis shows that ESPNet can learn meaningful representations for different medical imaging data efficiently.

Acknowledgment

Research reported in this chapter was supported by Washington State Department of Transportation research grant T1461-47, NSF III (1703166), National Cancer Institute awards (R01 CA172343, R01 CA140560, RO1 CA200690, and U01CA231782), the NSF IGERT Program (award no. 1258485), the Allen Distinguished Investigator Award, Samsung GRO award, and gifts from Google, Amazon, and Bloomberg. The authors would also like to thank Dr. Jamen Bartlett and Dr. Donald L. Weaver for helpful discussions and feedback. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing endorsements, either expressed or implied, of National Cancer Institute or the National Institutes of Health, or the US Government.

References

[1] C.J. Lynch and C. Liston, New machine-learning technologies for computer-aided diagnosis. s.l, *Nat. Med.* **24**, 2018, 1304–1305.

[2] E. Mercan, et al., Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions, *JAMA Netw. Open* **2**, 2019, e198777.

[3] LeCun, Y., Bengio, Y., Convolutional networks for images, speech, and time series. Volume The handbook of brain theory and neural networks. 1995.

[4] G. Litjens, et al., A survey on deep learning in medical image analysis, *Med. Image Anal.* **42**, 2017, 60–88.

[5] Shen, D., Wu, G., Suk, H.-I., Deep learning in medical image analysis. Annual review of biomedical engineering, 2017, pp. 221–248.

[6] Mehta, S. et al., Y-net: Joint segmentation and classification for diagnosis of breast biopsy images. s.l., International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018b, pp. 893–901.

[7] Kamnitsas, K. et al., Ensembles of multiple models and architectures for robust brain tumour segmentation. s.l., International MICCAI Brainlesion Workshop. 2017.

[8] Ronneberger, O., Fischer, P. & Brox, T., U-Net: Convolutional Networks for Biomedical Image Segmentation. s.l., International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.

[9] A. Esteva, et al., Dermatologist-level classification of skin cancer with deep neural networks, *Nature* **542**, 2017, 115–118.

[10] Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. s.l., International Conference on Medical Image Computing and Computer-Assisted Intervention. 2013.

[11] Mehta, S. et al., ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. s.l., Proceedings of the European Conference on Computer Vision (ECCV). 2018c.

[12] Mehta, S., Rastegari, M., Shapiro, L., Hajishirzi, H., ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network. s.l., IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[13] Nuechterlein, N., Mehta, S., 3D-ESPNet with Pyramidal Refinement for Volumetric Brain Tumor Image Segmentation. s.l., International MICCAI Brainlesion Workshop. 2018.

[14] L.-C. Chen, et al., Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Anal. Mach. Intell.* 2017, 834–848.

[15] Yu, F., Koltun, V., Multi-scale context aggregation by dilated convolutions. s.l., International Conference on Representation Learning. 2016.

[16] V. Badrinarayanan, A. Kendall and R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. pattern Anal. Mach. Intell.* 2017, 2481–2495.

[17] He, K., Zhang, X., Ren, S., Sun, J., Deep Residual Learning for Image Recognition. s.l., IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 770–778.

[18] Simonyan, K., Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition. s.l., International Conference on Representation Learning (ICLR). 2015.

[19] J.G. Elmore, et al., Diagnostic concordance among pathologists interpreting breast biopsy specimens, *JAMA* **313**, 2015, 1122–1132.

[20] J.G. Elmore, et al., A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis, *J. Pathol. Inform.* **8**, 2017, 12.

[21] Mehta, S. et al., Learning to segment breast biopsy whole slide images. s.l., 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018a, pp. 663–672.

[22] Zhao, H. et al., Pyramid scene parsing network. s.l., IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 2881–2890.

[23] B.H. Menze, et al., The multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Trans. Med. Imaging* **34**, 2015, 1993–2024.

[24] Myronenko, A., 3D MRI brain tumor segmentation usingautoencoder regularization. s.l., International MICCAI Brainlesion Workshop. 2018.

[25] Li, B. et al., A Facial Affect Analysis System for Autism Spectrum Disorder. s.l., IEEE International Conference on Image Processing (ICIP). 2019.

Footnotes

1 For simplicity, we assume that n is an odd number, such as 3, 5, 7, and so on.

2 Dilated convolutions are sometimes also referred to as *atrous convolutions*, wherein “trous” means holes in French.

3 Point-wise convolutions are a special case of standard $n \times n$ convolutions when $n = 1$. These convolutions produce an output plane by learning linear combinations between input channels. These convolutions are widely used for either channel expansion or channel reduction operations.

4 Gridding artifacts can be removed by adding another convolution layer with no dilation after the concatenation operation in Fig. 6.4 (left); however, this will increase the computational complexity of the block.

Abstract

Medical imaging is a fundamental part of clinical care that creates informative, noninvasive, and visual representations of the structure and function of the interior of the body. With advancements in technolog and the availability of massive amounts of imaging data, data-driven methods, such as machine learning and data mining, have become popular in medical imaging analysis. In particular, deep learning-based methods such as convolutional neural networks, now have the requisite volume of data and computational power to be considered practical clinical tools. We describe the architecture of the ESPNet network and provid experimental results for the task of semantic segmentation on two different types of medical images: (1) tissue-level segmentation of breast biopsy whole slide images and (2) 3D tumor segmentation in brain magneti resonance images. Our results show that the ESPNet architecture is efficient and learns meaningful representations for different types of medical images, which allows ESPNet to perform well on these images.

Keywords: ESPNet; whole slide images; magnetic resonance images; medical imaging; tumor segmentation

Queries and Answers

Query:

Please check all the author names and affiliations.

Answer: Yes, the author's names are correct.

Query:

Please provide city and country name for affiliation 3.

Answer: Change of affiliation 3. Please use the University of Washington instead of Allen Institute of Artificial intelligence and XNOR.AI

Query:

Please note that we have set keywords. Please check and amend if necessary.

Answer: Correct

Query:

Please provide expansion for PAP, FLAIR, and MNI.

Answer: PAP: **Papanicolaou test,** FLAIR: Fluid attenuated inversion recovery

Query:

Please note that we have introduced Fig. 6.9 citation here. Please check and amend if necessary.

Answer: Looks correct

Query:

Please alter the significance both in caption and image of Figs. 6.1–6.3, and 6.5 and image of Figs. 6.6 and 6.9 to ensure that they are meaningful when the chapter is reproduced both in color and B&W.

Answer: Check our comments