Dense Volume-to-Volume Vascular Boundary Detection

Jameson Merkow[1(✉)], Alison Marsden[2], David Kriegman[1], and Zhuowen Tu[1]

[1] University of California, San Diego, USA
jmerkow@eng.ucsd.edu
[2] Stanford University, Stanford, USA

Abstract. In this work, we tackle the important problem of dense 3D volume labeling in medical imaging. We start by introducing HED-3D, a 3D extension of the state-of-the-art 2D edge detector (HED). Next, we develop a novel 3D-Convolutional Neural Network (CNN) architecture, I2I-3D, that predicts boundary location in volumetric data. Our fine-to-fine, deeply supervised framework addresses three critical issues to 3D boundary detection: (1) efficient, holistic, end-to-end volumetric label training and prediction (2) precise voxel-level prediction to capture fine scale structures prevalent in medical data and (3) directed multi-scale, multi-level feature learning. We evaluate our approaches on a dataset consisting of 93 medical image volumes with a wide variety of anatomical regions and vascular structures. We show that our deep learning approaches out-perform the current state-of-the-art in 3D vascular boundary detection (structured forests 3D), by a large margin, as well as HED applied to slices. Prediction takes about one minute on a typical $512 \times 512 \times 512$ volume, when using GPU.

1 Introduction

The past decade has witnessed major progress in computer vision, graphics, and machine learning, due in large part to the success of technologies built around the concept of "image patches". Many patch-centric approaches fall into the category of "sliding-window" methods [3,9,11] that consider dense, overlapping windows. Patch-centric approaches limit us in terms of computational complexity and long-range modeling capabilities. Fully convolutional neural networks (FCN) [7] achieved simultaneous performance and full image labeling. Holistically-Nested Edge Detector (HED) [13] applied this approach to image-to-image object boundary detection. HED significantly improved the state-of-the-art in edge detection, and did so at a fraction of the computational cost of previous CNN-based edge/boundary detection algorithms. Another member of the FCN family, UNet [10], adapted this architecture for neuronal segmentation.

Volume-to-volume learning has yet to garner the same attention as image-to-image labeling. One approach applies 2D prediction schemes on images generated by traversing the volume on an anatomical plane then recombining predictions into a volume. However, volumetric features exist across three spatial dimensions, therefore it is crucial to process this data where those features exist.

S. Ourselin et al. (Eds.): MICCAI 2016, Part III, LNCS 9902, pp. 371–379, 2016.
DOI: 10.1007/978-3-319-46726-9_43

The current state-of-the-art in vessel wall detection uses a 3D patch-to-patch approach along with domain features, *a-priori* information, and a structured forest classifier [9]. In that work, the authors mitigate computational cost of patch-centric classifiers by using a sampling scheme and limiting their dataset to certain types of vascular structures. This method side-steps patch-centric inefficiency by limiting accurate prediction to a subset of structures and anatomical regions. Volumetric labeling using a CNN approach has been attempted [14], but the high computational cost of these frameworks preclude them from accurate end-to-end volumetric prediction.

A secondary challenge lies in detecting small structures prevalent in medical volume data. In contrast to objects in natural images, anatomical structures are often small, and resolution may be limited by acquisition. In fact, small anomalies are often of greater importance than the larger structures. These factors manifest a unique challenge for dense labeling of medical volumes.

In this work, we first extend HED (2D-CNN) to HED-3D for direct dense volume labeling; we then propose a novel 3D-CNN architecture, I2I-3D, for precise volume-to-volume labeling. Our approach tackles three key issues in dense medical volume label prediction: (1) efficient volumetric labeling of medial data using 3D, volume-to-volume CNN architectures, (2) precise fine-to-fine and volume-to-volume labeling, (3) nested multi-scale learning. We extend the typical fine-to-coarse architecture by adding an efficient means to process high resolution features late in the network, enabling precise voxel-level prediction that benefit from coarse level guidance and nested multi-scale representations. We evaluate our approach against the state-of-the-art in vessel wall detection.

2 Dense Volume-to-Volume Prediction

2.1 Pixel Level Prediction in 2D Images

Fully convolutional neural networks [7] were among the first methods to adapt the fine-to-coarse structure to dense pixel-level prediction. The FCN architecture added element-wise summations to VGGNet [2] that link coarse resolution predictions to layers with finer strides. However, it has been shown that pulling features directly from bottom layers to top layers is sub-optimal as the fine-level features have no coarse-level guidance [4]. HED [13] produced top accuracy on the BSDS500 dataset [8] with an alternative adaptation of VGGNet which fused several boundary responses at different resolutions with weighted aggregation. However, HED's fine-to-coarse framework leaves fundamental limitations to precise prediction and a close look at the edge responses produced by HED reveals many thick orphan edges. HED only achieves top accuracy after boundary refinement via non-maximum suppression (NMS) and morphological thinning. This approach is often sufficient for 2D tasks, however, it is less reliable in volumetric data. Furthermore, NMS fails when the prediction resolution is lower than the object separation resolution.

In these architectures, the most powerful outputs (in terms of predictive power) lack the capability to produce fine resolution predictions. Not only is

this problematic for making high resolution predictions, but coarse representations inform finer resolution predictions and finer resolution distinctions often require complex predictive power. UNet [10] addressed some of these issues by adding more convolutional layers, and using a loss function that penalizes poor localization of adjacent structures. However, UNet does not directly learn nested multi-level interactions and the large number of dense layers hinder efficiency in 3D tasks.

2.2 Precise Multi-Scale Voxel Level Prediction

Our framework addresses these crucial issues to volume-to-volume labeling and applies them to vascular boundary detection in medical volumes. Our proposed network, I2I-3D, consists of two paths: a fine-to-coarse path and a multi-scale coarse-to-fine path. The fine-to-coarse network structure follows popular network architecture and generates features with increasing feature abstraction and greater spatial extent. By adding side outputs and a fusion layer, we obtain an efficient 3D-CNN: HED-3D. As expected, HED-3D struggles to localize small vascular structures, and requires a secondary path to increases prediction resolution. I2I-3D uses a secondary path to learn complex multi-scale interactions in a coarse-to-fine fashion creating a fine-to-fine architecture.

Each stage of the coarse-to-fine path incorporates abstract representations with higher resolution features to produce fine resolution responses that benefit from multi-scale influences and coarse level guidance. Special 'mixing' layers and two convolution layers combine these two inputs to minimize a multi-scale, deeply supervised loss function. Here, deep supervision [6], plays an important role through multiple loss functions that reward multi-scale integration at each stage. Cascading this process results in features with large projective fields at high resolution. Later layers benefit from abstract features, coarse level guidance, and multi-scale integration; this culminates in a top most layer with the best predictive power and highest resolution. Figure 1 depicts the layer-wise connected 3D convolutional neural network architectures of I2I-3D and HED-3D.

2.3 Formulation

We denote our input training set of N volumes by $S = \{(X_n, Y_n), n = 1, \ldots, N\}$, where sample $X_n = \{x_j^{(n)}, j = 1, \ldots, |X_n|\}$ denotes the raw input volume and $Y_n = \{y_j^{(n)}, j = 1, \ldots, |X_n|\}, y_j^{(n)} \in \{1, .., K\}$ denotes the corresponding ground truth label map. For our task, $K = 2$, here we define the generic loss formulation. We drop n for simplicity, as we consider volumes independently. Our goal is to learn network parameters, $\mathbf{W}$, that enable boundary detection at multiple resolutions. Our approach produces M multi-scale outputs with $\frac{1}{2^{M-1}}$ input resolution. Each output has an associated classifier whose weights are denoted $\mathbf{w} = (\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(M)})$. Loss for each of these outputs is defined as:

$$\mathcal{L}_{\text{out}}(\mathbf{W}, \mathbf{w}) = \sum_{m=1}^{M} \ell_{\text{out}}^{(m)}(\mathbf{W}, \mathbf{w}^{(m)}), \tag{1}$$

where ℓ_{out} denotes the volume-level loss function. Loss is computed over all voxels in a training volume X and label map Y. Specifically, we define the following cross-entropy loss function used in Eq. (1):

$$\ell_{\text{out}}^{(m)}(\mathbf{W}, \mathbf{w}^{(m)}) = -\sum_{k} \sum_{j \in Y_k} \log \Pr(y_j = k | X; \mathbf{W}, \mathbf{w}^{(m)}) \tag{2}$$

where Y_k denotes the voxel truth label sets for the k^{th} class. $\Pr(y_j = k|X; \mathbf{W}, \mathbf{w}^{(m)}) = \sigma(a_j^{(m)}) \in [0, 1]$ is computed using sigmoid function $\sigma(.)$ on the activation value at voxel j. We obtain label map predictions $\hat{Y}_{\text{out}}^{(m)} = \sigma(\hat{A}_{\text{out}}^{(m)})$, where $\hat{A}_{\text{out}}^{(m)} \equiv \{a_j^{(m)},\ j = 1, \ldots, |Y|\}$ are activations of the output of layer m. Putting everything together, we minimize the following objective function via standard stochastic gradient descent:

$$(\mathbf{W}, \mathbf{w}) = \text{argmin}(\mathcal{L}_{\text{out}}(\mathbf{W}, \mathbf{w})) \tag{3}$$

During testing, given image X we obtain label map predictions from the output layers: $\hat{Y}_{\text{top}} = \text{I2I}(X, (\mathbf{W}, \mathbf{w}))$, where $\text{I2I}(\cdot)$ denotes the label maps produced by our network.

3 Network Architecture and Training

The coarse-to-fine path of I2I-3D, we mimic VGGNet's [2] design with domain specific modifications. First, we truncate at the fourth pooling layer resulting in a network with 10 convolutional layers at four resolutions. Second, we decrease the filter count of the first two convolution layers to 32. Lastly, we replace max pooling with average pooling. For our HED-3D framework, we place deep supervision at side-outputs at each convolution layer just prior to pooling, as in [13]. These side-outputs are fused via weighted aggregation.

I2I-3D adds a pathway to HED-3D's architectures to combines multi-scale responses into higher resolution representations. The second structure follows an inverted pattern of the fine-to-coarse path; it begins at the lowest resolution and upsamples in place of pooling. Each stage of the coarse-to-fine path contains a mixing layer and two convolutional layers. Mixing layers take two inputs: one from the corresponding resolution in fine-to-coarse path and a second from the output of the previous (coarser) stage in the coarse-to-fine path. Mixing layers concatenate inputs and do a specialized $1 \times 1 \times 1$ convolution operation to mix multi-resolution input features. Mixing layers are similar to reduction layers in GoogLeNet [12] but differ in usage and initialization. Mixing layers directly hybridize low resolution and fine-to-coarse features, while maintaining network efficiency. Mixing layer output is pass through two convolutional layers to spatially mix the two streams. Each coarse-to-fine stage is deeply supervised after the final convolution layer, just prior to upsampling. These side outputs push each stage to produce higher quality predictions by incorporating information from the lower resolution, more abstract representations These outputs are

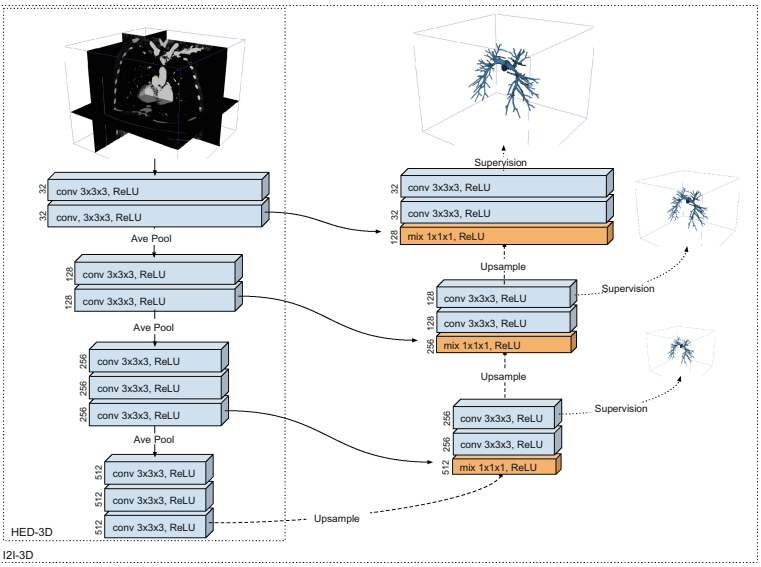


Fig. 1. The proposed network architecture I2I-3D. Our architecture couples fine-to-coarse and coarse-to-fine convolutional structures and multi-scale loss to produce dense voxel-level labels at input resolution. The number of channels is denoted on the left of each convolution layer, arrows denote network connections and operations.

only used to promote multi-scale integration at each stage, and are phased out, leaving single output at the top-most layer.

We begin by describing the training procedure for HED-3D. We, first, load pre-trained weights (details in Sect. 4) and place deep supervision at each of the four multi-scale outputs and at the fusion output. We iteratively train starting with a learning rate of $1e^{-7}$ and decimate every $30k$ iterations. Weight updates occur after every iteration till convergence.

For I2I-3D, we attach a coarse-to-fine path and move deep supervision to new multi-scale outputs. Each stage is initialize to produce the identity mapping of the fine-to-coarse input. We decrease all learning rates in the fine-to-coarse path to $\frac{1}{100}$ and train until loss plateaus. These hyper-parameters force the network to learn multi-scale features at each stage in order to minimize loss at each resolution. Finally, we return learning rate multipliers to 1, remove all supervision on all outputs except the highest resolution, and train until convergence.

4 Experimentation and Results

In [9], the authors use direct voxel overlap for evaluation, however, this metric fails to account for any localization error in boundary prediction and over-penalizes usable boundaries that do not perfectly overlap with ground truth boundaries. The metrics used here a 3D extension of the BSDS benchmark

metrics [1] which are standard protocols for evaluating boundary contours in natural images.

These metrics find match correspondences between ground truth and predicted contour boundaries. Matched voxels contribute to true positive counts, and unmatched voxels contribute to fall-out and miss rates. We report three performance measures: fixed threshold F measure (ODS), best per-image threshold F measure (OIS), and average precision (AP) and show precision-recall curves for each classifier.

Our dataset includes all 38 volumes used in [9] but introduces an 55 additional volumes to form an expanded dataset with 93 volumes. This dataset includes a variety of anatomical regions, including: abdominal, thoracic, cerebral, vertebral, and lower extremity regions. All volumes are accompanied by 3D models which were expertly built for computational blood flow simulation. Volumes were captured from individual patients for clinically indicated purposes via magnetic resonance (MR) or computed tomography (CT) imaging. Volumes include normal physiologies as well as a variety of pathologies including: aneurysms, stenoses, peripheral artery disease, congenital heart disease, and dissection. The dataset contains various arterial vessel types, but only one structure is annotated per volume. All volumes were obtained from http://www.vascularmodel.com.

We split volumes into training, validation, and test sets, each set contains 67, 7 and 19 volumes respectively and consist of both CT and MR data. Since volumes contain incomplete annotation, only voxels inside annotated vessels and those within 20 voxels of the vessel wall are considered during evaluation.

We pre-process each volume by whitening voxel intensities and cropping them into overlapping, $96 \times 96 \times 48$, segments. A single segment takes about one second to process on a NVidia K40 GPU and a typical volume ($512 \times 512 \times 512$) takes less than a minute. As a result of inconsistent annotation, we only train on volumes that contain over 0.25 % labeled vessel voxels (approx. 1000 of 442,368 voxels).

Our networks are implemented in the popular *Caffe* library [5] where methods were extended for 3D when necessary. Fine-to-coarse weights were generated by pre-training a randomly initialized network on entire vessel label prediction for a fixed number of iterations (50k) with a high learning rate. These labels produce less overall loss, preventing unstable gradients from developing during back-propagation.

We compare I2I-3D to the current state-of-the-art [9], a 2D-CNN baseline (HED) [13] and our HED-3D architecture. HED (in 2D) was trained without modification on individual slices from each volume. We also compare against the widely used 3D-Canny edge detector.

Figures 2 and 3 show the results of our experimentation. Figure 3 indicates that our method out-performs all other methods, including the current state-of-the-art. We also notice that 3D-CNN approaches considerably improves average precision over 2D-CNN when comparing results from HED and HED-3D. The precision-recall curves reveal that I2I-3D consistently boosts precision over HED-3D indicating that our fine-to-fine multi-scale architecture improves localization.

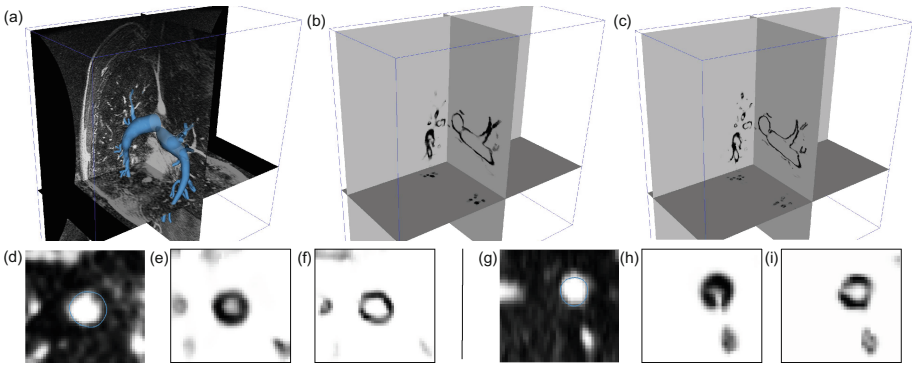
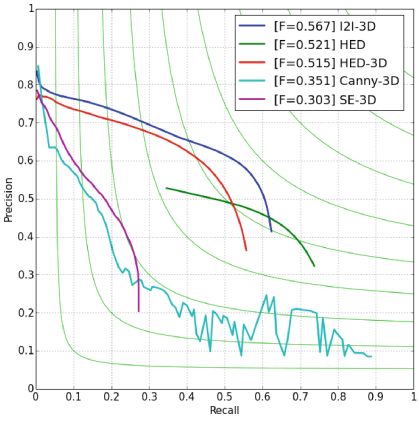


Fig. 2. Results of our HED-3D and I2I-3D vessel boundary classifiers. (a) Input volume and ground truth (in blue). (b) HED-3D result. (c) I2I-3D result. (d), (g) vessel cross section and ground truth (in blue). (e), (h) HED-3D cross section result. (f), (i) I2I-3D cross section result. (Color figure online)



	ODS	OIS	AP
SE-3D [9]	0.303	0.316	0.149
Canny-3D	0.351	0.545	0.241
HED [13]	0.521	0.542	0.182
HED-3D (ours)	0.515	0.528	0.362
I2I-3D (ours)	**0.567**	**0.580**	**0.421**

Fig. 3. (left) Precision recall curves comparing our approach with state-of-the-art, our baseline methods. (right) Performance metrics of our approach and baselines.

In Fig. 2, we see the results of I2I-3D characterized by stronger and more localized responses when compared to HED-3D, showing the benefit of our fine-to-fine, multi-scale learning approach. The fine-to-coarse architecture of HED and HED-3D generate low resolution responses resulting in poor localization of tiny structures and a weaker edge response. This indicates that I2I-3D's multi-scale representation enable precise localization.

5 Conclusion

We have proposed two network structures, HED-3D and I2I-3D, that address major issues in efficient volume-to-volume labeling. Our HED-3D framework

demonstrates that processing volumetric data natively in 3D, improves performance over its 2D counterpart; our framework, I2I-3D, efficiently learns multi-scale hierarchal features and generates precise voxel-level predictions at input resolution. We demonstrate through experimentation, that our approach is capable of fine localization and achieves state-of-the-art performance vessel boundary detection without explicit *a-priori* information. We provide our source code and pre-trained models to ensure that our approach can be applied to variety of medical applications and domains at: https://github.com/jmerkow/I2I.

Acknowledgments. Z.T. is supported by NSF IIS-1360566, NSF IIS-1360568, and a Northrop Grumman Contextual Robotics grant. We are grateful for the generous donation of the GPUs by NVIDIA.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. PAMI **33**(5), 898–916 (2011)
2. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: BMVC (2014)
3. Dollár, P., Zitnick, C.L.: Fast edge detection using structured forests. In: PAMI (2015)
4. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR (2014)
5. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. preprint arXiv:1408.5093 (2014)
6. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS (2015)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2014)
8. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. PAMI **26**(5), 530–549 (2004)
9. Merkow, J., Tu, Z., Kriegman, D., Marsden, A.: Structural edge detection for cardiovascular modeling. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 735–742. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24574-4_88
10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24574-4_28
11. Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M.: DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 556–564. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24553-9_68

12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
13. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
14. Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D.: 3D deep learning for efficient and robust landmark detection in volumetric data. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 565–572. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24553-9_69