Multi-Class Micro-CT Image Segmentation Using Sparse Regularized Deep Networks

Amirsaeed Yazdani*[1], Yung-Chen Sun*[1], Nicholas B. Stephens[2], Timothy Ryan[2], Vishal Monga[1]
[1] Department of Electrical Engineering, [2] Department of Anthropology, The Pennsylvania State University, University Park, Pennsylvania, USA.

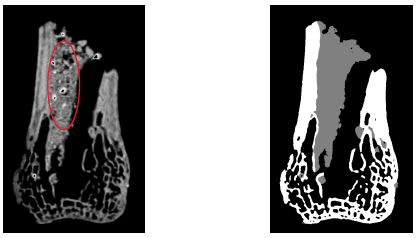
***Abstract*—It is common in anthropology and paleontology to address questions about extant and extinct species through the quantification of osteological features observable in micro-computed tomographic (μCT) scans. In cases where remains were buried, the grey values present in these scans may be classified as belonging to air, dirt, or bone. While various intensity-based methods have been proposed to segment scans into these classes, it is often the case that intensity values for dirt and bone are nearly indistinguishable. In these instances, scientists resort to laborious manual segmentation, which does not scale well in practice when a large number of scans are to be analyzed. Here we present a new domain-enriched network for three-class image segmentation, which utilizes the domain knowledge of experts familiar with manually segmenting bone and dirt structures. More precisely, our novel structure consists of two components: 1) a representation network trained on special samples based on newly designed custom loss terms, which extracts discriminative bone and dirt features, 2) and a segmentation network that leverages these extracted discriminative features. These two parts are jointly trained in order to optimize the segmentation performance. A comparison of our network to that of the current state-of-the-art U-NETs demonstrates the benefits of our proposal, particularly when the number of labeled training images are limited, which is invariably the case for μCT segmentation.**

***Index Terms*—Micro-CT, Osteology, Fossil, Quantitative morphometrics, Deep learning, Neural networks, U-NET**

I. Introduction

Micro-computed tomography (μCT) is increasingly used in palaeontology and anthropology to address fundamental questions about functional morphology, evolutionary history, and phylogenetic constraint [1]–[4]. Often these analyses rely on the quantitative analysis of bone, which is frequently the only remaining tissue that survives long-term burial after an animal dies. As such, their accuracy relies on a faithful segmentation of μCT scan intensity values belonging to air, bone, and non-bone material. In burial contexts, non-bone intensity values tend to belong to extraneous soils of varying densities, which are introduced following soft tissue decomposition. While the intensity values belonging to soil may differ dramatically from those of bone and air, there are many cases where heterogeneous densities form a continuum of values between bone and non-bone pixels, which limits the efficacy of intensity-based segmentation approaches [5]–[8]. In these instances, laborious manual segmentation is performed by an expert prior to analysis. An example of an unsegmented image and its corresponding ground truth can be seen in figure1.

*These authors contributed equally

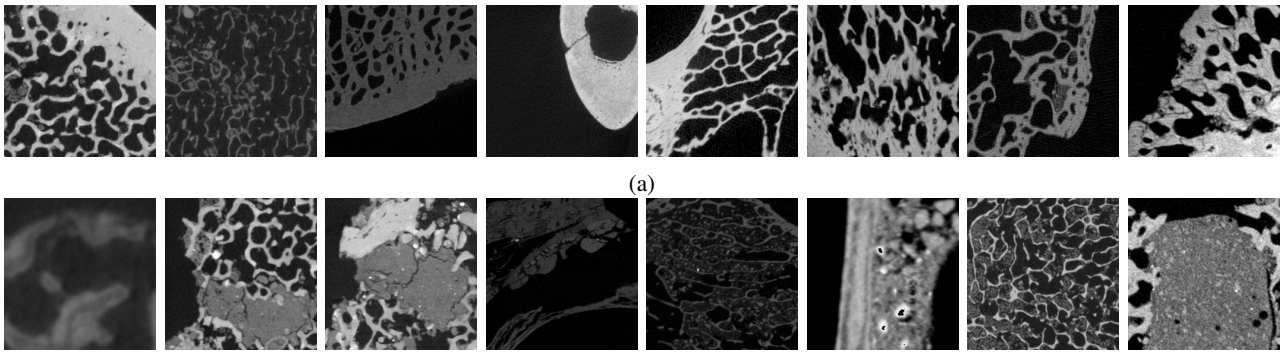


(a) Original (b) Ground truth

Fig. 1: A sample of an unsegmented image and its ground truth. The region inside the red circle looks like bone, however it is actually dirt (gray pixels in the ground truth)

While slice-by-slice manual segmentation is able to resolve many ambiguities that arise, it is untenable when a single high-resolution scan may result in a 3D volume with large grids (2040 x 2040) and thousands of tomographic slices. Sophisticated automatic approaches have been proposed, such as the k-means with fuzzy c-mean clustering algorithm [9], but intensity-based clustering results in occasional misclassification of non-bone material adjacent to bone. Alternative edge detection methods have also been applied to palaeontological material [10], but the implementation is not freely available. There are very few examples of machine learning applications to the segmentation of trabecular bone structure [11], which are limited by the quantity and quality of training data.

Deep learning frameworks applied to various imaging tasks have shown promising results, with image segmentation problems being of particular note [12]–[20]. In these cases, correspondence between unsegmented images and their segmented labels are mapped via sets of non-linear activation functions. Despite their success in segmentation tasks, their accuracy usually degrades when salient features are lacking in the underlying raw data. Solving this problem requires the incorporation of domain knowledge into the training dataset for the segmentation network to succeed. Previously, [20]–[23] developed networks that exploit special characteristics of segmentation problems using geometric priors. Even so, their prior guided segments only two classes, which for our problem would entail combining dirt and air into a general non-bone class. As depicted in figure 1, the similarities between bone and dirt intensity values heighten the risk of erroneously classifying soil as bone. However, despite the similarities between these two classes, they differ in terms of connectivity



(a)

(b)

Fig. 2: Examples of bone (a) and dirt (b) patches. Although in terms of pixel intensity the two groups look very similar, they have some discriminative features that the network tries to capture.

and distribution over all structures. In this work we propose a novel regularized network for three-class segmentation of bone, which consists of 1) custom representation layers where discriminative features of bone and dirt are extracted and 2) a segmentation network which uses the extracted features from the representation layers to segment the images. The representation layers have two blocks, one for each class, which are themselves trained using representative dirt and bone patches extracted from the dataset using custom regularization terms in addition to the segmentation loss. In particular, the regularizer for the 'dirt' class explicitly enforces sparsity, capturing the relatively small spatial footprint of dirt structures in the images. These patches are chosen such that the bone block has a high response to bone patches and a low response to dirt patches, and vice versa. We refer to this proposed network as a "Discriminative Sparse Regularized Deep Network for μ-CT segmentation" (DS-RDN).

II. Discriminative Sparse Regularized Deep Networks for μCT Segmentation

A. Background and Notation

Let $X \in \mathbf{R}^{M \times N}$ represent the input image. Let $Y \in \mathbb{R}^{M \times N}$ be the output segmented image and $Y_g \in \{0,1,2\}^{M \times N}$ be the manually-labeled y segmentation map corresponding to X where 0, 1, and 2 represent air, dirt, and bone, respectively. Note that this type of labeling essentially means that the outputs/labels for the network are vectors/matrices with three channels, in which each channel corresponds to each class and is set to one when a certain pixel belongs to that class (one-hot encoding). Let $W^B_{R_k}$ and $W^D_{R_k} \in \mathbb{R}^{m \times n}$ be the k^{th} filter in the geometric representation layer of bone and dirt, respectively, where m, n represent the width and height of the filter, respectively. Similarly, let $W^l_{S_k} \in \mathbb{R}^{m \times n \times d}$ be the k^{th} convolutional filter in layer l of the segmentation network, where m, n, and d represent the width, height, and depth of the filter, respectively. The representation and segmentation networks are, $\Theta_R = \{W^B_{R_k}, W^D_{R_k}\} \forall k$ and $\Theta_S = \{W^l_{S_k}\} \forall l, k$, respectively. Finally, let the mapping function of the representation network be represented by f, and the mapping function of the segmentation network be represented by F. Then, $Y = F(f(X, \Theta_R), \Theta_S)$. The network parameters Θ_R and Θ_S are learned by minimizing a loss function so as to produce an output Y that closely mimics the ground truth map Y_g. Since we are dealing with multi-class segmentation we use the cross entropy loss function where we have:

$$L(\Theta_S, \Theta_R) = -\sum_{i=1}^{3} Y_g^i \log(\tilde{Y}^i) \tag{1}$$

Where Y_g^i is the i-th channel of ground truth and $\tilde{Y}^i$ is the network output after being passed to the Sigmoid function. From a probabilistic point of view, we aim to make the distribution of the output as close as possible to the ground truth distribution for each class.

B. Representation Layer

The representation layers are trained jointly with the segmentation network to obtain two blocks specifically optimized for catching bone or dirt features. Feature extraction is complicated as a result of the shared intensity values and organization of dirt and bone (figure1). Therefore we define two blocks, each of which is in charge of detecting bone or dirt features. We define bone-dirt constraints for these blocks to grasp discriminative features and discard similarities. These constraints are imposed on the representation layer filters but they indirectly influence the segmentation network, since their joint optimization with the segmentation network propagates the changes.

C. Bone-Dirt Constraint

To obtain bone/dirt features, each of the bone/dirt blocks should be sensitive to bone/dirt patches and non-sensitive to dirt/bone patches. To achieve this, we define loss terms for each of the blocks so that they will be optimized with regard to their dedicated class features. Extractions of bone/dirt patches were carried out on challenging portions considered to be faithful representatives of each corresponding class. Figure 2 shows some of these patches for each class, where the bone patches in figure 2 a) and dirt patches in figure 2 b) are

very similar in terms of pixel intensity. Optimization of the bone/dirt blocks is achieved by defining:

$$L_{Bone}(\theta_R) = \sum_{n=1}^{K}\sum_{i=1}^{P}(\alpha||W_{R_n}^{Bone} \circledast I_d^i||_2^2 - \beta||W_{R_n}^{Bone} \circledast I_b^i||_2^2) \quad (2)$$

$$L_{Dirt}(\theta_R) = \sum_{n=1}^{K}\sum_{i=1}^{P}(\gamma||W_{R_n}^{Dirt} \circledast I_b^i||_2^2 - \sigma||W_{R_n}^{Dirt} \circledast I_d^i||_2^2 + \varsigma||W_{R_n}^{Dirt} \circledast I_d^i||_1) \quad (3)$$

Where I_d^i and I_b^i are the i^{th} dirt and bone patches, respectively, K is the total number of representation layer filters, with α, β, γ, σ, and ζ being positive regularization constants. The negative signs mean our goal is to maximize those terms while positive signs mean minimization objectives. We also add a sparsity term to aid in extracting features from the dirt block, because structures in the dirt class patterning occupy a relatively small part of the image, hence making it sparse.

D. Segmentation Network

The features extracted by the representation layer are concatenated and passed to a modified U-NET segmentation network [19], which is an architecture that has been shown to perform well in various classification and segmentation problems [17], [20]. The U-NET modifications are as follows: 1) the input channels are adjusted based on the number of channels from the representation layers and 2) the number of output channels is increased to three for generating one-hot outputs. So the complete loss function of the domain-enriched network would be:

$$L(\Theta_S, \Theta_R) = -\sum_{i=1}^{3} Y_g^i \log(\tilde{Y}^i) + \lambda_1 L_{Bone} + \lambda_2 L_{Dirt} \quad (4)$$

where λ_1 and λ_2 are regularizer constants. The parameters of the representation layer (θ_R) and segmentation network (θ_S) are optimized using the stochastic gradient descent [24], [25]. Since all of the terms in the loss function are differentiable with regard to network parameters, standard back-propagation is performed. The proposed domain-enriched network is illustrated in figure 3:

III. Experimental Evaluation

Datasets and Experimental Setup: Human, non-human, and fossil post-cranial osteological samples (n≃30) were µCT scanned (30-50µm voxel resolution) either at the Center for Quantitative Imaging, Pennsylvania State University using an ONMI-X HD600 or GE v|tome|x L300 (180 kV, 110 mA, 2800-4800 views), or at the Cambridge Biotomography Centre, University of Cambridge using a Nikon XTH 225 ST HRCT (125 kV, 135 mA, 1080 views). Ground truth images were hand-segmented with a digital tablet by an expert with n≃9 years experience. Networks were trained with 20 images and evaluated using a set of 13 images. Two additional experiments were performed to 1) evaluate the performance based on a lower training set of 10 images and b) to evaluate the robustness of the networks by randomly choosing the train/test splits.

Network Architecture: The domain bone/dirt blocks are composed of $K = 16$ 3×3 filters using a Light-UNet segmentation network as proposed in, which is obtained by compressing the U-NET layers [19]

TABLE I: Comparison between different methods

Method	Dice Overlap		
	Air	Dirt	Bone
MLP [11]	0.9890	0.2193	0.8526
U-net [19]	0.9902	0.3507	.9342
RDN-CS	0.9902	**0.3780**	**0.9354**

Metrics for Evaluation: Dice overlap (F1) is used as an evaluation metric for each class individually, with the definitions $F1 = \frac{2TP}{2TP+FP+FN}$, with TP, FP, and FN representing true positives, false positives, and false negatives, respectively with regard to each certain class.

Patch Extraction and Parameter Selection: Following from other deep learning frameworks [12], [13], [19], nearly 1200 256×256 patches were extracted from the original images by using sliding windows stride length set at 32. In any given dataset, the number of bone patches might not equal the number of dirt patches. Therefore, in every training epoch, we randomly selected an equal number of both types of patches. The regularization parameters inside the representation loss were 1e-5 and 0.5 for α, β, λ, and η, respectively. The batch size was set to 44 with 25 training epochs. The learning rate started at 10^{-4} and decreased with a drop factor of 0.1 for every 8 epochs. Training and testing were conducted on an NVIDIA Titan X GPU (12GB) with the PyTorch package [26].

Normal Set-up:Table I shows a comparison of Dice overlap results for the full training set of 20 images and 13 evaluation images for the proposed network (RDN-CS), a typical U-NET, and MLP. The higher value (specifically for the dirt class) shows the efficacy of including bone/dirt blocks in our method. Figure4 visualizes the differences in segmentation accuracy in two images using the proposed network and a typical U-NET. The incorrectly classified pixels across all classes are shown in red, with the proposed network having fewer incorrectly classified pixels in both images.

First Experiment: To determine how incorporating domain knowledge into our network affected performance at lower training sizes, we trained our network and a typical U-NET using 10, instead of 20, training images. Figure 5 shows the Dice overlap results for each class using the original evaluation set. These results demonstrate that the proposed method performs well despite the halving of the training sample. Specifically, while there is a substantial degradation of performance in the typical U-NET (especially for the dirt class), performance degradation of the DS-RDN is relatively minimal.

Second Experiment: In order to assess how robust the networks were regarding variation in the train/test splits, we performed 10 separate training sessions of our network and a typical U-NET, using 10 different train/test splits of 20 images. The results depicted in Figure 6 show the Gaussian distribution obtained by considering the mean and variance of the Dice

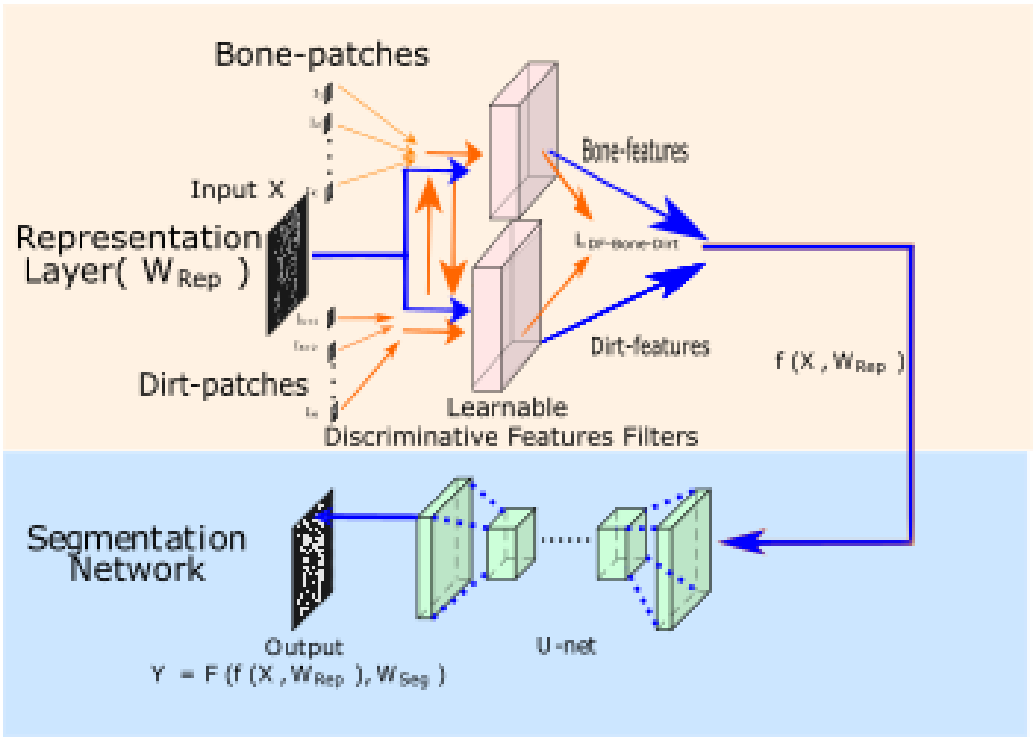


Fig. 3: The structure of the proposed network. The representation layer consists of two blocks, one each for bone and dirt. These blocks are trained jointly using the normal training images and extracted patches containing specific patterns, which are then used to compute the bone and dirt loss ($\boldsymbol{L_{Bone}}$ and $\boldsymbol{L_{Dirt}}$)

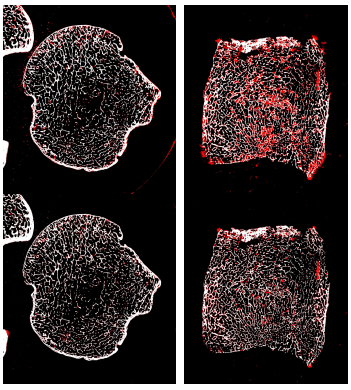


Fig. 4: The comparison of errors in two test samples segmented by a typical U-NET (top) and DS-RDN (bottom). Red pixels represent errors in segmentation across all classes.

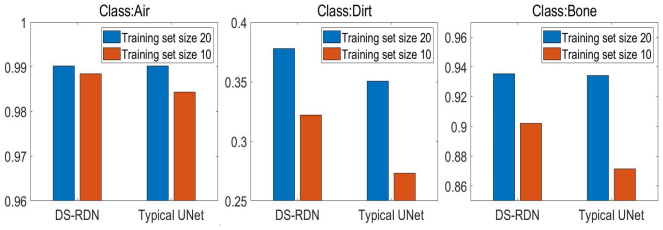


Fig. 5: Change in the Dice overlap in the first experiment as the training size decreases from 20 to 10 for DS-RDN and typical U-NET. From left to right: air, dirt and bone.

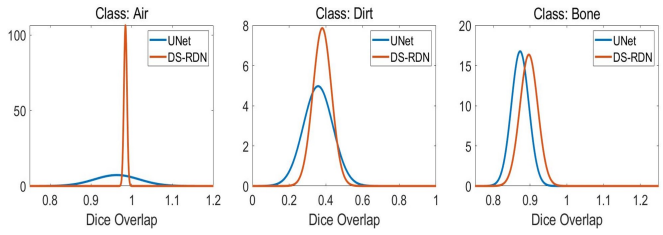


Fig. 6: The Gaussian distributions obtained by taking the mean and the variance of 10 different train/test splits (second experiment). From left to right: air, dirt, bone.

overlap over all splits for each architecture, with respect to each class. When compared to the distribution of the typical U-NET, the higher mean and narrow distribution of the proposed network demonstrates that it is more robust against variation in the train/test split for each class.

IV. Conclusion

Here we proposed a regularized custom network for multi-class segmentation of bone structure that incorporated domain knowledge to discriminate between two challenging classes. Multiple experimental designs were considered to measure the performance and stability of the proposed architecture versus well-established alternatives. When the training size was halved, we found the degradation of the proposed net-

work to be minimal compared to a typical U-NET. Similarly, we observed higher mean and narrower standard deviation distributions when varying the train/test data, indicating that the proposed network is more robust than a typical U-NET. Ultimately, these results support the incorporation of domain-specific knowledge in challenging segmentation tasks.

REFERENCES

[1] L. M. Witmer, R. C. Ridgely, D. L. Dufeau, and M. C. Semones, "Using CT to peer into the past: 3d visualization of the brain and ear regions of birds, crocodiles, and nonavian dinosaurs," in *Anatomical Imaging*. Springer Japan, 2008, pp. 67–87.

[2] C. A. Brassey, L. Margetts, A. C. Kitchener, P. J. Withers, P. L. Manning, and W. I. Sellers, "Finite element modelling versus classic beam theory: comparing methods for stress estimation in a morphologically diverse sample of vertebrate long bones," *Journal of The Royal Society Interface*, vol. 10, no. 79, pp. 20 120 823–20 120 823, nov 2012.

[3] M. M. Skinner, A. Evans, T. Smith, J. Jernvall, P. Tafforeau, K. Kupczik, A. J. Olejniczak, A. Rosas, J. Radovcic, J. F. Thackeray, M. Toussaint, and J. J. Hublin, "Brief communication: Contributions of enamel-dentine junction shape and enamel deposition to primate molar crown complexity," *Am J Phys Anthropol*, vol. 142, no. 1, pp. 157–63, May 2010, skinner, Matthew M Evans, Alistair Smith, Tanya Jernvall, Jukka Tafforeau, Paul Kupczik, Kornelius Olejniczak, Anthony J Rosas, Antonio Radovcic, Jakov Thackeray, J Francis Toussaint, Michel Hublin, Jean-Jacques eng Research Support, Non-U.S. Gov't Am J Phys Anthropol. 2010 May;142(1):157-63. doi: 10.1002/ajpa.21248. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20091839

[4] K. Harvati, C. Röding, A. M. Bosman, F. A. Karakostis, R. Grün, C. Stringer, P. Karkanas, N. C. Thompson, V. Koutoulidis, L. A. Moulopoulos, V. G. Gorgoulis, and M. Kouloukoussa, "Apidima cave fossils provide earliest evidence of homo sapiens in eurasia," *Nature*, vol. 571, no. 7766, pp. 500–504, 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-1376-z

[5] M. Doube, M. M. Kłosowski, I. Arganda-Carreras, F. P. Cordelières, R. P. Dougherty, J. S. Jackson, B. Schmid, J. R. Hutchinson, and S. J. Shefelbine, "Bonej: Free and extensible bone image analysis in imagej," *Bone*, vol. 47, no. 6, pp. 1076 – 1079, dec 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S8756328210014419

[6] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[7] A. Odgaard and H. Gundersen, "Quantification of connectivity in cancellous bone, with special emphasis on 3-d reconstructions," *Bone*, vol. 14, no. 2, pp. 173–182, mar 1993.

[8] T. W. Ridler and S. Calvard, "Picture thresholding using an iterative selection method," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 8, pp. 630–632, 1978.

[9] C. J. Dunmore, G. Wollny, and M. M. Skinner, "Mia-clustering: a novel method for segmentation of paleontological material," *PeerJ*, vol. 6, p. e4374, Feb. 2018. [Online]. Available: https://doi.org/10.7717/peerj.4374

[10] H. Scherf and R. Tilgner, "A new high-resolution computed tomography (CT) segmentation method for trabecular bone architectural analysis," *American Journal of Physical Anthropology*, vol. 140, no. 1, pp. 39–51, sep 2009.

[11] V. C. Korfiatis, S. Tassani, and G. K. Matsopoulos, "A new ensemble classification system for fracture zone prediction using imbalanced micro-ct bone morphometrical data," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1189–1196, July 2018.

[12] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2369–2380, 2016.

[13] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 109–118, Jan 2016.

[14] C. Cho, Y. H. Lee, and S. Lee, "Prostate detection and segmentation based on convolutional neural network and topological derivative," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3071–3074.

[15] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Išgum, "Automatic segmentation of mr brain images with a convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1252–1261, May 2016.

[16] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1562–1573, July 2018.

[17] P. Naylor, M. Laé, F. Reyal, and T. Walter, "Nuclei segmentation in histopathology images using deep neural networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, April 2017, pp. 933–936.

[18] Z. Yan, X. Yang, and K. Cheng, "Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1912–1923, Sep. 2018.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[20] A. Yazdani, N. B. Stephens, V. Cherukuri, T. Ryan, and V. Monga, "Domain-enriched deep network for micro-ct image segmentation," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 1867–1871.

[21] V. Cherukuri, V. K. B.G., R. Bala, and V. Monga, "Multi-scale regularized deep network for retinal vessel segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 824–828.

[22] V. Cherukuri, B. VijayKumar, R. Bala, and V. Monga, "Deep retinal image segmentation with regularization under geometric priors," *IEEE Transactions on Image Processing*, vol. 29, pp. 2552–2567, 2020.

[23] N. Stephens, A. Yazdani, V. Cherukuri, L. Demars, V. Monga, and T. Ryan, "Machine learning in anthropology: A regularized deep network for osteological micro-ct image segmentation," *American Journal of Physical Anthropology*, vol. In press., no. S69, 2020.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] P. J. Werbos, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. John Wiley & Sons, 1994.

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf