SPNet: Shape Prediction using a Fully Convolutional Neural Network

S M Masudur Rahman Al Arif[1], Karen Knapp[2] and Greg Slabaugh[1]

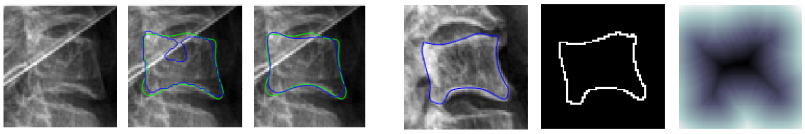
[1]City, University of London
[2]University of Exeter

Abstract. Shape has widely been used in medical image segmentation algorithms to constrain a segmented region to a class of learned shapes. Recent methods for object segmentation mostly use deep learning algorithms. The state-of-the-art deep segmentation networks are trained with loss functions defined in a pixel-wise manner, which is not suitable for learning topological shape information and constraining segmentation results. In this paper, we propose a novel shape predictor network for object segmentation. The proposed deep fully convolutional neural network learns to predict shapes instead of learning pixel-wise classification. We apply the novel shape predictor network to X-ray images of cervical vertebra where shape is of utmost importance. The proposed network is trained with a novel loss function that computes the error in the shape domain. Experimental results demonstrate the effectiveness of the proposed method to achieve state-of-the-art segmentation, with correct topology and accurate fitting that matches expert segmentation.

1 Introduction

Shape is a fundamental topic in medical image computing and particularly important for segmentation of known objects in images. Shape has been widely used in segmentation methods, like the statistical shape model (SSM) [1] and level set methods [2], to constrain a segmentation result to a class of learned shapes. Recently proposed deep fully convolutional neural networks show excellent performance in segmentation tasks [3,4]. However, the neural networks are trained with a pixel-wise loss function, which fails to learn high-level topological shape information and often fails to constrain the object segmentation results to possible shapes (see Fig. 1a, 1b and 1c). Incorporating shape information in deep segmentation networks is a difficult challenge.

In [6], a deep Boltzmann machine (DBM) has been used to learn a shape prior from a training set. The trained DBM is then used in a variational framework to perform object segmentation. A multi-network approach for incorporating shape information with the segmentation results was proposed in [7]. It uses a convolutional network to localize the segmentation object, an autoencoder to infer the shape of the object, and finally uses deformable models, a version of SSM, to achieve segmentation of the target object. Another method for localization of shapes using a deep network is proposed in [8] where the final segmentation is performed using SSM. All these methods consist of multiple components which are not trained in an end-to-end fashion and thus cannot fully utilize the excellent representation learning capability of neural networks



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

Fig. 1: (a-c) Advantage of shape prediction over pixel-wise classification (a) a noisy test image (b) segmentation result from a state-of-the-art deep network [5] (c) predicted shape from the proposed shape predictor network, SPNet. The green curve (—) represents the manually annotated vertebral boundary and the blue curve (—) represents the vertebral boundary of the predicted vertebra. The proposed SPNet can constrain the predicted shape to resemble a vertebra-like structure where the pixel-wise classification network failed in the presence of a strong image artifact. (d-f) Examples of a training vertebra (d) original image with manually annotated vertebral boundaries (e) pixels at the zero-level set (f) signed distance function. Darker tone represents negative values.

for shape prediction. Recently, two methods were proposed which utilize a single network to achieve shape-aware segmentation. The method proposed in [9] uses a shallow convolutional network which is trained in two-stages. First, the network is trained in a supervised manner. Then the network is fine-tuned by using unlabelled data where the ground truth are generated with the help of a level set-based method. In contrast, the work presented in [5], proposed a shape-based loss term for training a deep segmentation network. However, both of these methods still use a cross-entropy loss function which is defined in a pixel-wise manner and thus not suitable to learn high-level topological shape information and constraints. In contrast to these methods, we propose a novel deep fully convolutional neural network, that is able to predict shapes instead of classifying each pixel separately. To the best of our knowledge, this is the first work that uses a fully convolutional deep neural network for shape prediction. We apply the proposed shape predictor network for segmentation of cervical vertebra in X-ray images where shape is of utmost importance and has constrained variation limits.

Most of the work in vertebra segmentation involves shape prediction [10, 11]. Given the fact that a vertebra in an X-ray image mostly consists of homogeneous and noisy image regions separated by edges, active shape model and level set-based methods can be used to evolve a shape to achieve a segmentation [1, 2, 12]. While these methods work relatively well in many medical imaging modalities, inconsistent vertebral edges and lack of a difference in image intensities inside and outside the vertebra limits the performance of these methods in clinical X-ray image datasets.

Our proposed network is closely related to the state-of-the-art work on cervical vertebrae [5, 13]. As mentioned earlier, [5] proposed a shape-based term in the loss function for training a segmentation network, UNet-S. The modified UNet [3] architecture produces a segmentation map for the input image patch which is defined over the same pixel space as the input. The UNet was further modified in [13], to achieve probabilistic spatial regression (PSR). Instead of classifying each pixel, the PSR network was trained to predict a spatially distributed probability map localizing vertebral corners.

In this work, we modify this UNet architecture to generate a signed distance function (SDF) from the input image. The predicted SDF is converted to shape parameters compactly represented in a shape space, in which the loss is computed. The contribu-

tions of this paper are two-fold: we propose 1) an innovative deep fully convolutional neural network that predicts shapes instead of segmentation maps and 2) a novel loss function that computes the error directly in the shape domain in contrast to the other deep networks where errors are computed in a pixel-wise manner. We demonstrate that the proposed approach outperforms the state-of-the-art method with topologically correct results, particularly on more challenging cases.

2 Dataset and Ground Truth Generation

This work utilizes the same dataset of lateral cervical X-ray images used in [5, 13]. The dataset consists of 124 training images and 172 test images containing 586 and 797 cervical vertebrae, respectively. The dataset is collected from hospital emergency rooms and is full of challenging cases. The vertebra samples include low image intensity, high noise, occlusion, artifacts, clinical conditions such as osteophytes, degenerative change, and bone implants. The vertebral boundary of each vertebra in the dataset is manually annotated by expert radiologists (blue curve in Fig. 1d). The training vertebra patches were augmented using multiple scales and orientation angles. A total of 26,370 image patches are used for training the proposed deep network. The manual annotation for each of the training vertebrae is converted into a signed distance function (SDF). To convert the vertebral shapes into an SDF (Φ), the pixels lying on the manually annotated vertebral boundary curve have been assigned zero values. Then all other pixels are assigned values based on the infimum of the Euclidean distances between the corresponding pixel and the set of pixels with zero values. Mathematical details can be found in the supplementary materials. An example of the training vertebra with corresponding zero-level set pixels and SDF are illustrated in Fig. 1d, 1e and 1f. After converting all the training vertebral shapes to corresponding signed distance functions, principal component analysis (PCA) is applied. PCA allows each SDF (Φ) in the training data to be represented by a mean SDF ($\bar{\Phi}$), matrix of eigenvectors (W) and a vector of shape parameters, $\boldsymbol{b}$:

$$\boldsymbol{\phi} = \bar{\boldsymbol{\phi}} + W\boldsymbol{b}, \tag{1}$$

where $\boldsymbol{\phi}$ and $\bar{\boldsymbol{\phi}}$ are the vectorized form of Φ and $\bar{\Phi}$, respectively. For each training example, we can compute $\boldsymbol{b}$ as:

$$\boldsymbol{b} = W^T(\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}) = W^T\boldsymbol{\phi}_d, \tag{2}$$

where $\boldsymbol{\phi}_d$ is the vectorized difference SDF, $\Phi_d = \Phi - \bar{\Phi}$. These parameters are used as the ground truth ($\boldsymbol{b}^{GT}$) for training the proposed network.

3 Methodology

To choose an appropriate network architecture for the application in hand, we follow the state-of-the-art work on cervical vertebrae [5, 13]. We note that the choice can be altered based on the application, the complexity of the model and the available memory in the system for training. Our proposed shape predictor network, SPNet, takes a 64×64 vertebral image patch as input and produces its related difference SDF ($\hat{\Phi}_d$) which is also defined over the same pixel space. We use the same network architecture as [13]. However, the final normalization layer has been removed. Instead, the last convolution

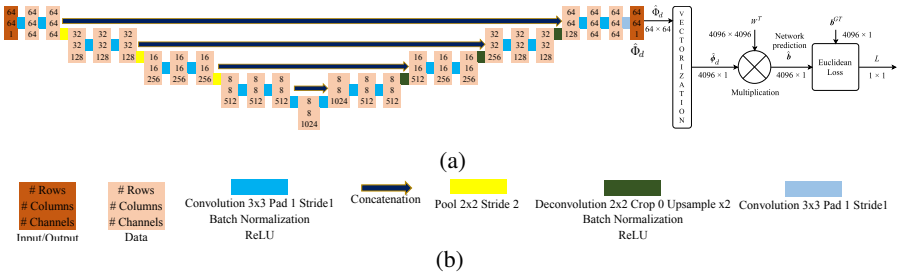


Fig. 2: SPNet: shape predictor network (a) network architecture (b) legend.

layer outputs the difference signed distance function ($\hat{\Phi}_d$) which is then sent to the final layer where it is converted to shape parameter vector ($\hat{\boldsymbol{b}}$) and compared with the ground truth ($\boldsymbol{b}^{GT}$). The network is illustrated in Fig. 2.

The forward pass through the final layer can be summarized below. First, the output of the last convolutional layer of the SPNet ($\hat{\Phi}_d$) is vectorized as $\hat{\boldsymbol{\phi}}_d$. Then the final prediction of network is computed as $\hat{\boldsymbol{b}}$:

$$\hat{\boldsymbol{b}} = W^T \hat{\boldsymbol{\phi}}_d \text{ or in the element-wise form: } \hat{b}_i = \sum_{j=1}^{k} w_{ij} \hat{\phi}_{d_j}, i = 1, 2, \cdots, k; \quad (3)$$

where w_{ij} is the value at the i-th row and j-th column of the transposed eigenvector matrix (W^T) and k is the number of shape parameters. Finally, the loss is defined as:

$$L = \sum_{i=1}^{k} L_i \text{ where } L_i = \frac{1}{2}(\hat{b_i} - b_i^{GT})^2. \quad (4)$$

The predicted shape parameter vector, $\hat{\boldsymbol{b}}$, has the same length as $\hat{\boldsymbol{\phi}}_d$ which is $64 \times 64 = 4096$. The initial version of the proposed network is designed to generate the full length shape parameter vector. However, the final version of the network is trained to generate fewer parameters which will be discussed in Sec. 5.

4 Experiments

The proposed network (SPNet) has been trained on a system with an NVIDIA Pascal Titan X GPU[1] for 30 epochs with a batch-size of 50 images. The network took approximately 22 hours to train. We have also implemented a traditional convolutional neural network (CNN) where we predict the shape parameter vector $\boldsymbol{b}$ directly using a Euclidean loss function. The network consists of the contracting path of the proposed SPNet architecture, followed by two fully connected (FC) layers which regress the 4096 b-parameters at the output. This network will be mentioned as SP-FCNet in the following discussions. The SPNet has only 24,237,633 parameters where the SP-FCNet network has 110,123,968 trainable parameters. The FC layers cause a significant increase in the number of parameters. For comparison, we also show results of vertebral shape prediction based on the Chan-Vese level set segmentation method (LS-CV) [2,14]. Apart from

[1] We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

these, we also compare our results with the segmentation networks described in [5]. Following their conventions, the shape-aware network will be referred to as UNet-S and the non-shape-aware version as UNet. The foreground predictions of these networks have been converted into shapes by tracking the boundary pixels. For the shape predictor networks, SPNet and SP-FCNet, the predicted b-parameters are converted into a signed distance function using Eqn. 1. The final shape is then found by locating the zero-level set of this function. We compare the predicted shapes with the ground truth shapes using two error metrics: the average point to ground truth curve error (E_{p2c}) and the Hausdorff distance (d_H) between the prediction and ground truth shapes. Both metrics are reported in pixels.

5 Results

We first compare the three shape prediction methods in Table 1. We report the mean and standard deviation of the metrics over 797 test vertebrae. The Chan-Vese method (LS-CV) achieves an average E_{p2c} of 3.11 pixels, whereas the fully connected version of the shape predictor network (SP-FCNet) achieves 2.27 pixels and the proposed UNet-based shape predictor network (SPNet) achieves only 1.16 pixels. Hausdorff distance (d_H) shows more difference between the LS-CV and the deep networks. The comparison also illustrates how the proposed SPNet is superior to its traditional CNN-based counterpart, SP-FCNet. Both of these networks predict the shape parameter vector ($\hat{\boldsymbol{b}}$) and the final loss is computed using Euclidean distance. It is the proposed SPNet's capabilities of generating the difference SDF ($\hat{\Phi}_d$) and backpropagating the Euclidean loss on the SDF (Eqn. 4) that make it perform better.

Table 1: Comparison of shape prediction methods.

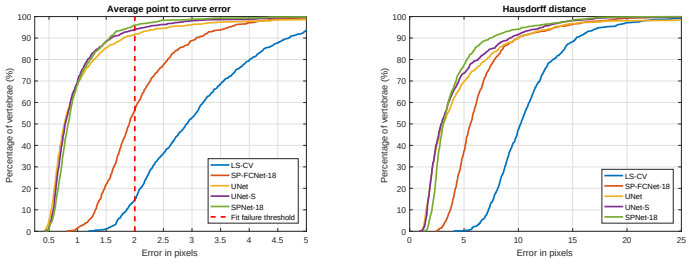
Metrics	Average E_{p2c} (pixel)		Average d_H (pixel)	
Methods	Mean	Std	Mean	Std
LS-CV	3.11	1.13	10.94	3.68
SP-FCNet	2.27	0.83	6.74	3.25
SPNet (proposed)	**1.16**	**0.66**	**4.11**	**3.13**

Both of the deep networks have been trained to regress all 4096 shape parameters which are related to the corresponding eigenvectors. As the eigenvectors are ranked based on their eigenvalues, eigenvectors with small eigenvalues often result from noise and can be ignored. We evaluated the trained SPNet on a validation set at test time by varying the number of predicted parameters. The best performance was observed when only the first 18 b-parameters are kept which represents 98% of the total variation in the training dataset.

Based on this insight, we modified both versions of our deep networks to regress only 18 b-parameters and retrained the networks from randomly initialized weights. We report the performance of the retrained networks in Table 2. We also report the metrics for UNet and UNet-S networks from [5]. It can be seen that our proposed SPNet-18, outperforms all other networks quantitatively. However, the improvement over UNet-S in terms of the E_{p2c} metric is small and not statistically significant according to the paired t-test at a 5% significance level. Quantitative improvements for SPNet-18 over all other cases pass the significance test.

Table 2: Quantitative comparison of different methods.

Metrics	Average E_{p2c} (pixel)		Average d_H (pixel)		$nVmR$	Fit failure (FF) %
Methods	Mean	Std	Mean	Std		
LS-CV	3.107	1.13	10.94	3.68	**0**	85.45
SP-FCNet-18	2.082	0.78	6.48	3.32	**0**	43.54
UNet	1.114	1.29	5.06	6.11	57	8.53
UNet-S	0.999	0.67	4.37	4.02	45	6.02
SPNet-18	**0.996**	**0.55**	**4.17**	**3.06**	**0**	**4.14**



(a) (b)

Fig. 3: Cumulative error curves (a) average E_{p2c} and (b) average d_H.

The most important benefit of the proposed SPNet over the UNet and UNet-S is that the loss is computed in the shape domain, not in a pixel-wise manner. In the fifth column of the Table 2, we report the **n**umber of test **v**ertebrae with **m**ultiple disjoint predicted **r**egions ($nVmR$). The pixel-wise loss function-based networks learn the vertebral shape implicitly, but this does not prevent multiple disjoint predictions for a single vertebra. The UNet and UNet-S produce 57 and 45 vertebrae, respectively with multiple predicted regions, whereas the proposed network does not have any such example indicating that the topological shape information has been learned based on the seen shapes. A few examples of these can be found in Fig. 4. We have also reported the fit failure (FF) for all the compared methods. Similar to [5], the FF is defined as the percentage of the test vertebrae having an E_{p2c} of greater than 2 pixels. The proposed SPNet-18 achieves the lowest FF. The cumulative error curves of the metrics are shown in Fig. 3. The performance of the proposed method is very close with the UNet and UNet-S in terms of the E_{p2c} metric. But in terms of the Hausdorff distance (d_H), the proposed method achieves noticeable improvement.

Moreover, the qualitative results in Fig. 4 distinctively demonstrate the benefit of using the proposed method. The UNet and UNet-S predict a binary mask and the predicted shape is located by tracking the boundary pixels. This is why the shapes are not smooth. In contrast, the proposed SPNet predicts b-parameters which are then converted to signed distance functions. The shape is then located based on the zero-level set of this function, resulting in smooth vertebral boundaries defined to the sub-pixel level which resembles the manually annotated vertebral boundary curves.

The worst performance is exhibited by the Chan-Vese method, LS-CV. The results of SP-FCNet-18 are better than the traditional Chan-Vese model, but underperform compared to the UNet-based methods. The reason can be attributed to the loss of spatial information because of the pooling operations. The UNet-based methods recover the spatial information in the expanding path by using concatenated data from the con-

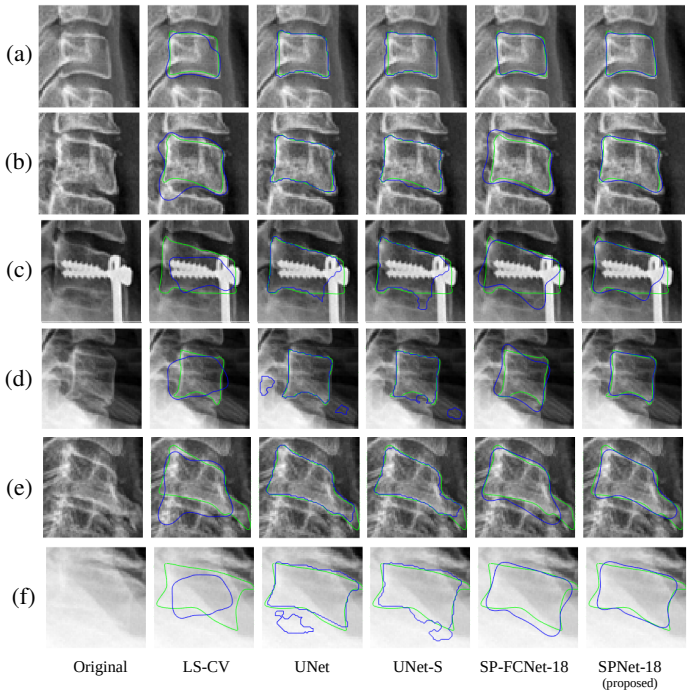


Fig. 4: Qualitative results: predicted shape (—) and ground truth (—).

tracting path, thus perform much better than the fully connected version of the deep networks. Some relatively easy examples are shown in Fig. 4a and 4b. More challenging examples with bone implants (Fig. 4c), abrupt contrast change (Fig. 4d), clinical condition (Fig. 4e) and low contrast (Fig. 4f) are also reported. It can be seen even in these difficult situations, the SPNet-18 method predicts shapes which resembles a vertebra where the pixel-wise loss function-based UNet and UNet-S predict shapes with unnatural variations. More qualitative examples and further results with a fully automatic patch extraction process are illustrated in the supplementary material, demonstrating our method's capability of adjusting to variations in scale, orientation, and translation of the vertebral patch.

6 Conclusion

In this paper, we have proposed a novel method which exploits the excellent representation learning capability of the deep networks and the pixel-to-pixel mapping capability of the UNet-like encoder-decoder architectures to generate object shapes from the input images. Unlike the pixel-wise loss function-based segmentation networks, the loss for the shape predictor network is computed in the shape parameter space. This encourages better learning of high-level topological shape information and restricts the predicted shapes to a class of training shapes.

The proposed shape predictor network can also be adapted for segmentation of other organs in medical images where preservation of the shape is important. The network proposed in this paper is trained for segmentation of a single object in the input image.

However, the level set method used for ground truth generation is inherently capable of representing object shapes that go through topological changes. Thus, given an appropriate object dataset, the same network can be used for segmentation of multiple and a variable number of objects in the input image. Similarly, the level set method can also be used to represent 3D object shapes. By replacing the UNet-like 2D deep network with a VNet-like [4] 3D network, our proposed method can be extended for 3D shape predictions. In future work, we plan to investigate the performance of our shape predictor network for segmentation of multiple and 3D objects.

References

1. T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995. 1, 2
2. A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. E. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Transactions on Medical Imaging*, vol. 22, no. 2, pp. 137–154, 2003. 1, 2, 5
3. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, Springer, 2015. 1, 2
4. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016. 1, 8
5. S. M. M. R. Al-Arif, K. Knapp, and G. Slabaugh, "Shape-aware deep convolutional neural network for vertebrae segmentation," in *5th International Workshop and Challenge in Computational Methods and Clinical Applications for Musculoskeletal Imaging (MICCAI MSKI)*, pp. 12–24, Springer, 2018. 2, 3, 5, 6
6. F. Chen, H. Yu, R. Hu, and X. Zeng, "Deep learning shape priors for object segmentation," in *Computer Vision and Pattern Recognition*, pp. 1870–1877, IEEE, 2013. 1
7. M. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri," *Medical Image Analysis*, vol. 30, pp. 108–119, 2016. 1
8. A. Mansoor, J. J. Cerrolaza, R. Idrees, E. Biggs, M. A. Alsharid, R. A. Avery, and M. G. Linguraru, "Deep learning guided partitioned shape model for anterior visual pathway segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1856–1865, 2016. 1
9. M. Tang, S. Valipour, Z. V. Zhang, D. Cobzas, and M. Jagersand, "A deep level set method for image segmentation," in *3rd International Workshop on Deep Learning in Medical Image Analysis (MICCAI DLMIA)*, vol. 10553, p. 126, Springer, 2017. 2
10. M. G. Roberts, T. F. Cootes, and J. E. Adams, "Automatic location of vertebrae on DXA images using random forest regression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 361–368, Springer, 2012. 2
11. S. M. M. R. Al-Arif, M. Gundry, K. Knapp, and G. Slabaugh, "Improving an active shape model with random classification forest for segmentation of cervical vertebrae," in *4th International Workshop and Challenge in Computational Methods and Clinical Applications for Spine Imaging (MICCAI CSI)*, vol. 10182, p. 3, Springer, 2017. 2
12. T. F. Chen, "Medical image segmentation using level sets," *Technical Report. Canada, University of Waterloo*, pp. 1–8, 2008. 2
13. S. M. M. R. Al-Arif, K. Knapp, and G. Slabaugh, "Probabilistic spatial regression using a deep fully convolutional neural network," in *British Machine Vision Conference, BMVC 2017, London, UK, September 4-7, 2017*, 2017. 2, 3
14. T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001. 5