# Deep Image-to-Image Recurrent Network with Shape Basis Learning for Automatic Vertebra Labeling in Large-Scale 3D CT Volumes

Dong Yang[1], Tao Xiong[2], Daguang Xu[3(✉)], S. Kevin Zhou[3], Zhoubing Xu[3], Mingqing Chen[3], JinHyeong Park[3], Sasa Grbic[3], Trac D. Tran[2], Sang Peter Chin[2], Dimitris Metaxas[1], and Dorin Comaniciu[3]

[1] Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA
[2] Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218, USA
[3] Medical Imaging Technologies, Siemens Healthcare Technology Center, Princeton, NJ 08540, USA
{daguang.xu,shaohua.zhou,sasa.grbic, dorin.comaniciu}@siemens-healthineers.com

**Abstract.** Automatic vertebra localization and identification in 3D medical images plays an important role in many clinical tasks, including pathological diagnosis, surgical planning and postoperative assessment. In this paper, we propose an automatic and efficient algorithm to localize and label the vertebra centroids in 3D CT volumes. First, a deep image-to-image network (DI2IN) is deployed to initialize vertebra locations, employing the convolutional encoder-decoder architecture. Next, the centroid probability maps from DI2IN are modeled as a sequence according to the spatial relationship of vertebrae, and evolved with the convolutional long short-term memory (ConvLSTM) model. Finally, the landmark positions are further refined and regularized by another neural network with a learned shape basis. The whole pipeline can be conducted in the end-to-end manner. The proposed method outperforms other state-of-the-art methods on a public database of 302 spine CT volumes with various pathologies. To further boost the performance and validate that large labeled training data can benefit the deep learning algorithms, we leverage the knowledge of additional 1000 3D CT volumes from different patients. Our experimental results show that training with a large database improves the performance of proposed framework by a large margin and achieves an identification rate of 89%.

## 1 Introduction

Accurate and automatic localization and identification of human vertebrae have become of great importance in 3D spinal imaging for clinical tasks such as pathological diagnosis, surgical planning and post-operative assessment of pathologies.

D. Yang and T. Xiong—Authors contributed equally.


M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 498–506, 2017.
DOI: 10.1007/978-3-319-66179-7_57

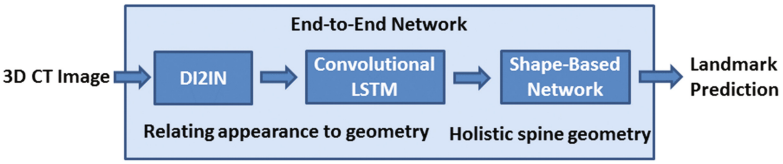


**Fig. 1.** Proposed method consisting of three major components: DI2IN, ConvLSTM and shape-based Network.

Specific applications such as vertebrae segmentation, fracture detection, tumor detection and localization, registration and statistical shape analysis can benefit from the efficient and precise vertebrae detection and labeling algorithms. However, designing such an algorithm requires addressing various challenges such as pathological cases, image artifacts and limited field-of-view (FOV).

In the past, many approaches have been developed to address these limitations in spine detection problems. In [1], Glocker *et al.* presented a two-stage approach for localization and identification of vertebrae in CT, which has achieved an identification rate of 81%. This approach uses a regression forests and a generative model for prediction and it requires handcrafted feature vectors in pre-processing. Then, Glocker *et al.* [2] further extended the vertebrae localization to handle pathological spine CT. This supervised classification forests based approach achieves an identification rate of 70% and outperforms state-of-the art on a pathological database. Recently, deep convolutional neural network has also been highlighted in the research of human vertebrae detection. A joint learning model with deep neural networks (J-CNN) [3] has been designed to effectively identify the type of vertebra and improved the identification rate (85%) with a large margin. They trained a random forest classifier to coarsely detect the vertebral centroids instead of directly performing neural network on the whole CT volumes. Suzani *et al.* [4] also presented a deep neural network for fast vertebrae detection. This approach first extracts the intensity-based features; then uses a deep neural network to localize the vertebrae. Although this approach has achieved high detection rate, it suffers from the large mean error compared to other approaches.

To meet the requirements of both accuracy and efficiency and take advantage of deep neural networks, we present an approach, shown in Fig. 1, with following contributions: *(a) Deep Image-to-Image Network (DI2IN) for Voxel-Wise Regression:* Instead of extracting handcrafted features or adopting coarse classifiers, the proposed deep image-to-image network directly performs on the 3D CT volumes and outputs the multichannel probability maps associated with different vertebrae centers. The high responses in probability maps intuitively indicate the location and label of vertebrae. The training is formulated as a multichannel voxel-wise regression. Since the DI2IN is implemented in a fully convolutional way, it is significantly efficient in time compared to the sliding-window approaches. *(b) Response Enhancement with ConvLSTM:* Inspired by [5], we introduce a recurrent neural network (RNN) to model the spatial

relationship of vertebra responses from DI2IN. The vertebrae can be interpreted in a chain structure from head to hip according to their related positions. The sequential order of the chain-structured model enables the vertebra responses to communicate with each other using recurrent model, such as RNN. The popular architecture, ConvLSTM, is adopted as our RNN to capture the spatial correlation between vertebra prediction. The ConvLSTM studies the pair-wise relation of vertebra responses and regularize the output of the DI2IN. *(c) Refinement using a Shape Basis Network:* To further refine the coordinates of vertebrae, we incorporate a shape basis network which takes advantage of the holistic structure of spine. Instead of learning a quadratic regression model to fit the spinal shape, we adopt the coordinates of spines in training samples to construct a shape-based dictionary and formulate the training process as a regression problem. The shape-based neural network extracts the coordinates from the previous stage as input and generates the coefficients associated with the dictionary, which indicates the linear combination of atoms from the shape-based dictionary. By embedding the shape regularity in the training of neural network, ambiguous coordinates are removed and the representation is optimized, which further improves the localization and identification performance. Compared to previous method [3] which applies classic refinement method as a post-processing step, our algorithm introduces an end-to-end training network in the refinement step for the first time, which allows us to train each component separately and then fine-tuned together in an end-to-end manner.

## 2 Method

### 2.1 Deep Image-to-Image Network (DI2IN)

In this section, we present the architecture and details of the proposed deep image-to-image network, as shown in Fig. 2. The basic architecture is designed as a convolutional encoder-decoder network [6]. Compared to sliding-window approach, the DI2IN is implemented in a voxel-wise fully convolutional end-to-end learning. It performs the network on 3D CT volumes directly. Basically, the DI2IN takes the 3D CT volume as input and generates the multichannel probability maps simultaneously. The ground truth probability maps are generated by Gaussian distribution $I_{gt} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\|\mathbf{x}-\mu\|^2/2\sigma^2}$, where $\mathbf{x} \in \mathbb{R}^3$ and $\mu$ denote the voxel coordinates and ground truth location, respectively. $\sigma$ is predefined to control the scale of the Gaussian distribution. Each channel's prediction $I_{prediction}$ is associated with the centroid location and type of vertebra. The loss function is defined as $|I_{prediction} - I_{gt}|^2$ for each voxel. Therefore, the whole learning problem is formulated as a multichannel voxel-wise regression. Instead of using classification formulation for detection, regression is tremendously helpful for determining predicted coordinates and it relieves the issue of imbalanced training samples, which is very common in semantic segmentation.

The encoder is composed of convolution, max-pooling and rectified linear unit (ReLU) layers while the decoder is composed of convolution, ReLU and upsampling layers. Max-pooling layers are of great importance to increase receptive field

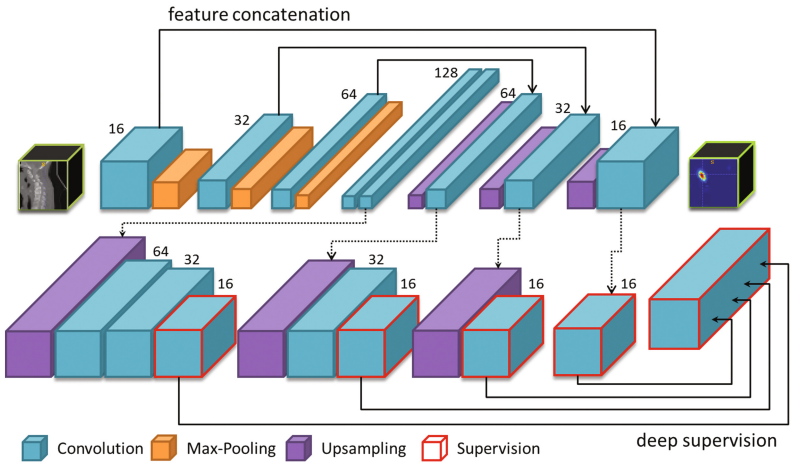


**Fig. 2.** Proposed deep image-to-image network (DI2IN). The front part is a convolutional encoder-decoder network with feature concatenation, and the backend is multi-level deep supervision network. Numbers next to convolutional layers are channel numbers. Extra 26-channel convolution layers are implicitly used in deep supervision.

and extract large contextual information. Upsampling layers are designed with the bilinear interpolation to enlarge and densify the activation, which also further enables the end-to-end voxel-wise training without losing resolution details. The convolutional filter size is $1 \times 1 \times 1$ in the output layer and $3 \times 3 \times 3$ in other layers. The max-pooling filter size is $2 \times 2 \times 2$ for down-sampling by half in each dimension. In upsampling layers, the input features are upsampled by a factor of 2 in each dimension. The stride is set as 1 in order to maintain the same size in each channel. Additionally, we incorporate the feature concatenation and deep supervision in DI2IN. In feature concatenation, a bridge is built directly from the encoder layer to the decoder layer, which passes forward the feature information from the encoder and then concatenates it with the decoder layer [7]. As a result, the DI2IN benefits from both local and global contextual information. Deep supervision has been adopted in [8–10] to achieve good boundary detection and organ segmentation. In the DI2IN, we incorporated a more complex deep supervision approach to further improve the performance. Several branches are diverged from the middle layers of the decoder network. With the appropriate upsampling and convolutional operations, the output size of all branches matches the size of 26-channel ground truth. In order to take advantage of deep supervision, the total loss function $loss_{total}$ of DI2IN is defined as the combination of loss $l_i$ for all output branches as follows.

$$loss_{total} = \sum_i loss_i + loss_{final} \tag{1}$$

### 2.2 Response Enhancement Using Multi-layer ConvLSTM

Given the image $I$, the DI2IN generates a probability map $P(v_i|I)$ for the centroid of each vertebra $i$ with high confidence. The vertebrae are localized at the peak positions $v_i$ of probability maps. However, we find that these probability maps are not perfect yet: some probability maps don't have response or have very low response at the ground truth locations because of similar image appearances of several vertebrae (e.g. $T1 \sim T12$). In order to handle the problem of missing response, we propose a RNN to effectively enhance the probability maps by incorporating prior knowledge of the spinal structure.

RNN has been widely developed and used in many applications, such as natural language processing, video analysis. It is capable to handle arbitrary sequences of input, and performs the same processing on every element of the sequence with memory of the previous computation. In our case, the spatial relation of vertebrae naturally forms a chain structure from top to bottom. Each element of the chain is the response map of a vertebra centroid. The proposed RNN model treats the chain as a sequence and enables vertebra responses of DI2IN to communicate with each other. In order to adjust the 3D response maps of vertebrae, we apply the convolutional LSTM (ConvLSTM) as our RNN model shown in Fig. 3. Because the $z$ direction is the most informative dimension, the $x, y$ dimensions are set to 1 for all the convolution kernels. During inference, we pass information forward and backward to regularize the output of DI2IN. The passing process can be conducted $k$ iterations ($k = 2$ in our experiments). All input-to-hidden and hidden-to-hidden operations are convolution. Therefore, the response distributions can be adjusted with necessary displacement or enhanced by the neighbors' responses.

Equation (2) describes how the LSTM unit is updated at each time step. $X_1, X_2, \ldots$ and $X_t$ are input states for vertebrae, cell states are $C_1, C_2, \ldots$ and $C_t$, and the hidden states are $H_1, H_2, \ldots$ and $H_t$. $i_t, f_t$ and $o_t$ are the gates of

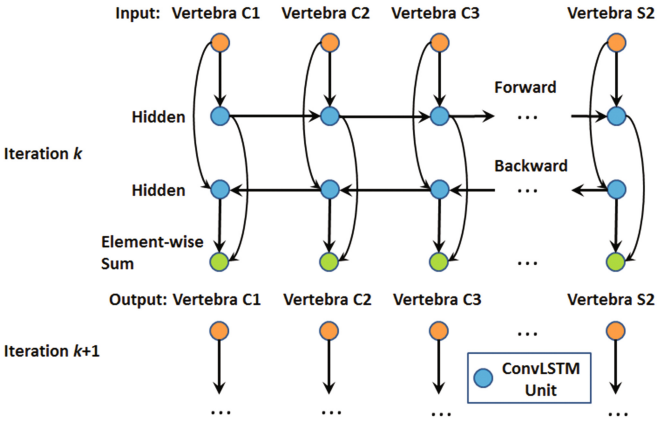


**Fig. 3.** The multi-layer ConvLSTM architecture for updating the vertebra response.

ConvLSTM. We use several sub-networks $G$ to update $X_t$, and $H_t$, which differs from the original ConvLSTM setting (original work only uses single kernel). Each $G$ is consist of three convolutional layers with $1 \times 1 \times 9$ kernels, and filter numbers are 9, 1 and 1. The sub-networks are more flexible and have a larger receptive field compared to that uses a single kernel. Therefore, it is helpful to capture the spatial relationship of all vertebrae.

$$\begin{aligned} i_t &= \sigma\left(G_{xi}\left(X_t\right) + G_{hi}\left(H_{t-1}\right) + W_{ci} \odot C_{t-1} + b_i\right) \\ f_t &= \sigma\left(G_{xf}\left(X_t\right) + G_{hf}\left(H_{t-1}\right) + W_{cf} \odot C_{t-1} + b_f\right) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tanh\left(G_{xc}\left(X_t\right) + G_{hc}\left(H_{t-1}\right) + b_c\right) \\ o_t &= \sigma\left(G_{xo}\left(X_t\right) + G_{ho}\left(H_{t-1}\right) + W_{co} \odot C_t + b_o\right)) \\ H_t &= o_t \odot \tanh\left(C_t\right) \end{aligned} \tag{2}$$

### 2.3 Shape Basis Network for Refinement

As shown in Fig. 4, the ConvLSTM generates clear probability maps, where the high response in the map indicates the potential location of the landmark (centroid of the vertebrae). However, sometimes due to image artifacts and low image resolution, it is difficult to guarantee there is no false positive. Therefore, we present a shape basis network to help refine the coordinates inspired by [11].

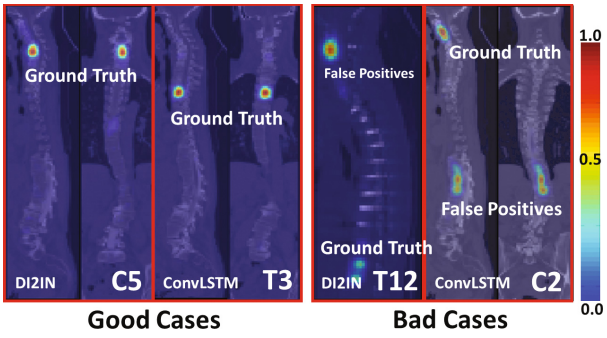


**Fig. 4.** Probability map examples from DI2IN (left in each case) and ConvLSTM (right in each case). The prediction in "Good Cases" is close to ground truth location. In "Bad Cases", some false positives exist remotely besides the response at the ground truth location.

Given a pre-defined shape-based dictionary $\mathbf{D} \in \mathbb{R}^{N \times M}$ and coordinate vector $\mathbf{y} \in \mathbb{R}^N$ generated by ConvLSTM, the proposed shape basis network takes $\mathbf{y}$ as input and outputs the coefficient vector $\mathbf{x} \in \mathbb{R}^M$ associated with dictionary $\mathbf{D}$. Therefore, the refined coordinate vector $\hat{\mathbf{y}}$ is defined as $\mathbf{Dx}$. In practice, the shape-based dictionary $\mathbf{D}$ is simply learned from the training samples. For example, the dictionary $\mathbf{D}_z$ associated with the vertical axis is constructed by

the $z$ coordinate of vertebrae centroids in the training sample. $N$ and $M$ indicate the number of vertebrae and number of atoms in dictionary, respectively.

The proposed shape basis network consists of several fully connected layers. Instead of regressing the refined coordinates, the network is trained to regress the coefficients $\mathbf{x}$ associated with the shape-based dictionary $\mathbf{D}$. The learning problem is formulated as a regression model and the loss function is defined as:

$$loss_{shape} = \sum_{i} ||\mathbf{D}\mathbf{x}_i - \mathbf{y}_i||_2^2 + \lambda||\mathbf{x_i}||_1 \tag{3}$$

$\mathbf{x}_i$ and $\mathbf{y}_i$ denote the coefficient vector and ground truth coordinate vector of $i$th training sample. $\lambda$ is the $\ell_1$ norm coefficient to leverage the sparsity and residual. Intuitively, the shape-based neural network is learned to find out the best linear combination in the dictionary to refine the coordinates. In our case, we focus on the refinement of vertical coordinates. The input of shape basis network is obtained directly from the output of ConvLSTM using a non-trainable fully connected layer. The layer has uniform weights and no bias term, and it generates the correct coordinates when the response is clear. Such setting enables the end-to-end scheme for fast inference instead of solving the loss function directly.

## 3 Experiments

First, we evaluate the proposed method on database introduced in [2] which consists of 302 CT scans with various types of lesions. The dataset has some cases with unusual appearance, such as abnormal spinal structure and bright visual artifacts due to metal implants by post-operative procedures. Furthermore, the FOV of each CT image varies greatly in terms of vertical cropping, image noise and physical resolution [1]. Most cases contain only part of the entire spine. The overall spinal structure can be seen only in a few examples. Large changes in lesions and limited FOV increase the complexity of the appearance of the vertebrae. It is difficult to accurately localize and identify the spinal column. The ground truth is marked on the center of gravity of each vertebra and annotated by the clinical experts. In previous work [1,3,4], two different settings have been conducted on this database: the first one uses 112 images as training and other 112 images as testing. The second one takes all data in first setting plus extra 18 images as the training data (overall 242 training images), and 60 unseen images are used as the testing data. For fair comparison, we follow the same configuration, which are referred as Set 1 and Set 2 respectively, in the experiments. Table 1 compares our result with the numerical results reported in previous methods [2–4] in terms of the Euclidean distance error (mm) and identification rate (Id.Rates) defined by [1]. The average mean errors of these two databases are 10.6 mm and 8.7 mm, respectively, and the identification rates are 78% and 85%, respectively. Overall, the proposed method is superior to the state-of-the-art methods on the same database with respect to mean error and identification rate.

**Table 1.** Comparison of localization errors in *mm* and identification rates among different methods. Our method is trained and tested using default data setting in "Set 1" and "Set 2", while "+1000" indicates training with additional 1000 labeled spine data and evaluated on the same testing data.

| Region | Method | Set 1 | | | Set 2 | | |
|---|---|---|---|---|---|---|---|
| | | Mean | Std | Id.Rates | Mean | Std | Id.Rates |
| All | Glocker *et al.* [2] | 12.4 | 11.2 | 70% | 13.2 | 17.8 | 74% |
| | Suzani *et al* [4] | 18.2 | 11.4 | - | - | - | - |
| | Chen *et al.* [3] | - | - | - | 8.8 | 13.0 | 84% |
| | Our method | 10.6 | **8.7** | 78% | 8.7 | 8.5 | 85% |
| | Our method +1000 | **9.0** | 8.8 | **83%** | **6.9** | **7.6** | **89%** |

We collect additional 1000 CT volumes and train the proposed DI2IN from scratch to verify whether training a neural network with more labeled data will improve its performance. This data set covers large visual changes of the spinal column (e.g. age, abnormality, FOV, contrast etc.). We evaluated on the same database and reported the results in Table 1 (shown as "Our method + 1000 training data"). As can be seen, adding more training data will greatly improve the performance of the proposed method, verifying that a large amount of labeled data will effectively boost the power of DI2IN.

## 4 Conclusion

In this paper, we presented an accurate and automatic method for human vertebrae localization and identification in 3D CT volumes. Our approach outperformed other state-of-the-art methods of spine detection and labeling in terms of localization mean error and identification rate.

**Acknowledgements.** We thank Dr. David Liu who provided insight and expertise that greatly assisted the research.

## References

1. Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E.: Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7512, pp. 590–598. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33454-2_73
2. Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A.: Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 262–270. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40763-5_33

3. Chen, H., Shen, C., Qin, J., Ni, D., Shi, L., Cheng, J.C.Y., Heng, P.-A.: Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 515–522. Springer, Cham (2015). doi:10.1007/978-3-319-24553-9_63
4. Suzani, A., Seitel, A., Liu, Y., Fels, S., Rohling, R.N., Abolmaesumi, P.: Fast automatic vertebrae detection and localization in pathological CT scans - a deep learning approach. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 678–686. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_81
5. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using CNNs. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 230–238. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_27
6. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 (2015)
7. 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, Cham (2015). doi:10.1007/978-3-319-24574-4_28
8. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403 (2015)
9. Merkow, J., Kriegman, D., Marsden, A., Tu, Z.: Dense volume-to-volume vascular boundary detection. arXiv preprint arXiv:1605.08401 (2016)
10. Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.-A.: 3D deeply supervised network for automatic liver segmentation from CT volumes. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 149–157. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_18
11. Yu, X., Zhou, F., Chandraker, M.: Deep deformation network for object landmark localization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 52–70. Springer, Cham (2016). doi:10.1007/978-3-319-46454-1_4