Deep Fusion Net for Multi-Atlas Segmentation: Application to Cardiac MR Images

Heran Yang[1], Jian Sun[1], Huibin Li[1], Lisheng Wang[2], and Zongben Xu[1]

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, China
[2] Department of Automation, Shanghai Jiaotong University, China

Abstract. Atlas selection and label fusion are two major challenges in multi-atlas segmentation. In this paper, we propose a novel deep fusion net for better solving these challenges. Deep fusion net is a deep architecture by concatenating a feature extraction subnet and a non-local patch-based label fusion (NL-PLF) subnet in a single network. This network is trained end-to-end for automatically learning deep features achieving optimal performance in a NL-PLF framework. The learned deep features are further utilized in defining a similarity measure for atlas selection. Experimental results on Cardiac MR images for left ventricular segmentation demonstrate that our approach is effective both in atlas selection and multi-atlas label fusion, and achieves state of the art in performance.

Keywords: Multi-atlas segmentation, deep fusion net, feature learning, atlas selection, end-to-end training, left ventricular segmentation.

1 Introduction

Multi-atlas segmentation (MAS) aims at segmenting anatomical structures or tissues from a target image by fusing the ground-truth segmentation labels of multiple atlases [6]. It has been one of the most popular methodologies over the past decade. In MAS, atlas images are warped to the target image by registration, and then corresponding warped atlas labels are combined to produce an estimated segmentation of the target image, *i.e.*, *target label*.

Atlas selection and *label fusion* are two major steps in multi-atlas segmentation. *Atlas selection* is to select a few most relevant atlases for a target image, so as to raise computational efficiency or improve final segmentation accuracy. It relies on a ranking of atlases, and several similarity measures between atlas and target image have been proposed [4], [9]. *Label fusion* is to predict the target label by fusing the warped atlas labels, and a key problem is the accurate computation of fusion weights for atlas pixels or patches. Non-local patch-based label fusion (NL-PLF) approach [1, 2] has been the state of the art in MAS, which uses all the patches in a search region around a pixel of interest for label fusion. Besides, [1], [5] extract hand-crafted features to compute the fusion weights, while [11] utilizes dictionary learning for label fusion.

In this work, we propose a novel deep architecture for multi-atlas segmentation, dubbed *deep fusion net*, which comprises a *feature extraction subnet* for feature extraction, and a *non-local patch-based label fusion subnet* for label fusion.

Deep fusion net can be interpreted as a deep architecture for feature learning in NL-PLF framework. Compared to NL-PLF methods using features extracted by handcraft or unsupervised learning, we discriminatively learn optimal deep features for label fusion by concatenating feature extraction and NL-PLF in a single network structure. Moreover, we apply the extracted features to define an atlas distance for atlas selection, shown to be effective in experimental section.

We test our method on a cardiac MR image set provided by Cardiac Atlas Project in MICCAI 2013 SATA Segmentation Challenge [1]. The data were collected from patients with coronary artery diseases and regional wall motion abnormalities due to prior myocardial infarction. The experiments demonstrate that deep fusion net can effectively select well-aligned atlases and accurately fuse atlas labels. Compared to the traditional methods, our proposed method achieves state of the art in Dice metric for left ventricular segmentation.

Deep learning has been applied to multi-atlas organ segmentation [3], [8]. These methods commonly learn a classification net as a pixel-wise or segment-wise label predictor. Contrary to them, our net is a label fusion net relying on the registration of atlas to target image. We learn deep features to compute optimal fusion weights for fusing the warped atlas labels provided by registration. This reduces the ambiguities in classification purely based on local patches. The advantage of fusion net compared to a classification net is shown in section 3.2.

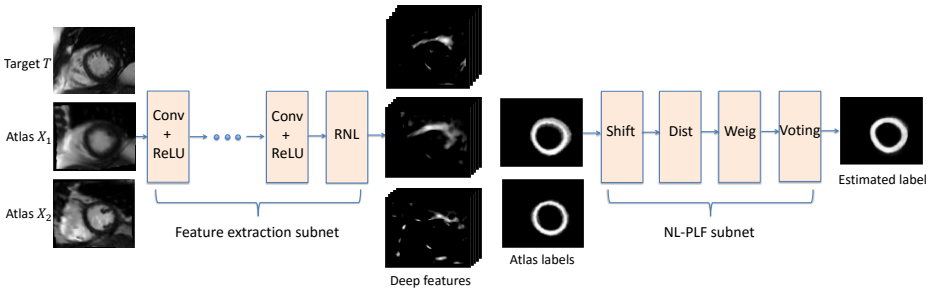


Fig. 1. The architecture of deep fusion net. A target image T and its warped atlas images X_i are first fed into a *feature extraction subnet* to extract deep features. These features and atlas labels are then sent to a *non-local patch-based label fusion subnet (NL-PLF subnet)* for generating the estimated label of target image, *i.e.*, target label.

2 Method: Deep Fusion Net

As shown in Fig. 1, *deep fusion net* (DFN) for multi-atlas segmentation is defined as a *feature extraction subnet*, followed by a *non-local patch-based label fusion subnet (NL-PLF subnet)*. The first subnet is responsible for extracting dense features from target and atlas images, and the second one aims at fusing the warped atlas labels using these extracted features. Given a target image T, we

[1] `https://masi.vuse.vanderbilt.edu/workshop2013/index.php`

register multiple atlases to T, and the warped atlas image and label pairs are denoted as $\{X_i, L(X_i)\}_{i=1}^K$. *First*, target image T and atlas images $\{X_i\}_{i=1}^K$ are fed into feature extraction subnet to output their per-pixel features $F(T)$ and $\{F(X_i)\}_{i=1}^K$. *Then*, these features and atlas labels are fed into NL-PLF subnet to generate target label. The network parameters are learned by end-to-end training based on a loss defined between network output and ground-truth target label.

2.1 Feature Extraction Subnet

The feature extraction subnet extracts deep features from images. All the target images and warped atlas images share the same feature extraction subnet. As shown in Fig. 1, the subnet consists of multiple repetitions of convolutional layer with ReLU activation function, and a final layer for response normalization.

Convolutional layer [7] convolves input feature using a set of learnable filters $\{\mathcal{W}^k\}_{k=1}^{D'}$, and each filter $\mathcal{W}^k \in \mathbb{R}^{w_f \times w_f \times D}$ is a third-order tensor. Given an input feature $G^{l-1}(X) \in \mathbb{R}^{M \times N \times D}$ of image X, this layer outputs feature $G^l(X) \in \mathbb{R}^{(M-w_f+1)\times(N-w_f+1)\times D'}$ at layer l. Rectified linear unit (ReLU) function is defined as $\varphi(x) = \max(0, x)$.

Response normalization layer (RNL) [7] normalizes feature for robustly computing feature distance in NL-PLF subnet. Given an input feature $G(X)$ for image X with element $g_{m,n,d}$, the normalized feature is $F(X)$ with element: $f_{m,n,d} = g_{m,n,d} / \left(\kappa + \alpha \sum_{i=max(0,d-\beta/2)}^{min(D-1,d+\beta/2)} (g_{m,n,i})^2\right)^{\gamma}$, where D is the size of the third dimension of $G(X)$. As in [7], κ, α, β and γ are set to 2, 10^{-4}, 5 and 0.75.

2.2 Non-Local Patch-Based Label Fusion Subnet

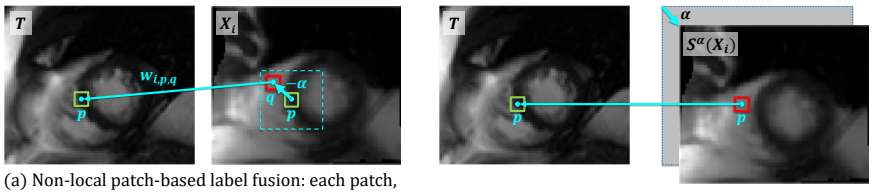
This subnet is a deep architecture implementing non-local patch-based label fusion scheme on top of feature extraction subnet. As shown in Fig.1, our NL-PLF subnet consists of *shift layer, distance layer, weight layer* and *voting layer*, and outputs the estimated label of target image.

Figure 2(a) shows the idea of non-local patch-based label fusion scheme. For each pixel p in a target image T, all the atlas labels in non-local search window around p in the warped atlases $\{X_i\}_{i=1}^K$ are fused to estimate the target pixel label. In deep fusion net, fusion weights are computed using the deep features extracted by the feature extraction subnet. Specifically, the fusion weight of pixel q in atlas X_i for predicting the label of pixel p in target T is computed by

$$w_{i,p,q}(\Theta) = \frac{\exp(-||F_p(T;\Theta) - F_q(X_i;\Theta)||^2)}{\sum_j \sum_{q \in N_p} \exp(-||F_p(T;\Theta) - F_q(X_j;\Theta)||^2)}, \tag{1}$$

where Θ is the network parameters, *i.e.*, filters and biases, in feature extraction subnet. $F_p(T;\Theta)$ is the extracted feature vector of image T at pixel p, N_p is a search window around p. Hence, the estimated label of p in target image T is

$$\hat{L}_p(T;\Theta) = \sum_i \sum_{q \in N_p} w_{i,p,q}(\Theta) L_q(X_i), \tag{2}$$



(a) Non-local patch-based label fusion: each patch, *e.g.,* patch around q, in a window around p votes for the label of pixel p in T

(b) Shift operation: Feature distance between p of T and q of X_i in (a) equals to feature distance of T and $S^\alpha(X_i)$ at p

Fig. 2. Illustration of non-local patch-based label fusion and shift operation.

where $L_q(X_i)$ is the label of X_i at pixel q.

We expect that the estimated label of T using Eqn. (2) should be close to the ground-truth label $L(T)$. Therefore, we define a *loss layer*:

$$E(\hat{L}(T;\Theta), L(T)) = \frac{1}{P}\left\|\hat{L}(T;\Theta) - L(T)\right\|^2, \tag{3}$$

where P is the number of pixels in $L(T)$. Our task in network training is to minimize this loss function on a training set *w.r.t.* the network parameters Θ using back-propagation. We next omit Θ for brevity.

Directly computing fusion weights using Eqn.(1) is fairly complex in gradient computation, because it is highly non-linear and depends on the pairwise feature distances in search windows. Hence, we decompose this complex operation into successive simple operations modeled as *shift layer*, *distance layer* and *weight layer*. Each operation and the gradient of its output *w.r.t* input can be efficiently computed using GPU in network training.

Figure 2 shows our motivation for this decomposition. We observe that, the feature distance of pixel p in T and q in X_i shown in Fig. 2(a) equals to the per-pixel feature distance at p between T and the shifted X_i in direction $\alpha = p - q$ shown in Fig. 2(b). Suppose that the search window width is $2t + 1$. To compute Eqn.(1), we first shift each feature map of X_i in each direction $\alpha \in R_{nl} = \left\{(u, v) \in \mathbb{Z}^2 | -t \leq u, v \leq t\right\}$ using a *shift layer*, then compute the pixel-wise feature distance using a *distance layer*, and finally transform these distances into voting weights using a *weight layer*.

Shift Layer. It spatially shifts the feature or label of an atlas. For feature $F(X_i) \in \mathbb{R}^{M\times N\times D}$ of atlas X_i, this layer generates $(2t+1)\times(2t+1)$ spatially shifted features along each direction $\alpha \in R_{nl}$. Due to the boundary effect, we remove the boundary and only retain features of pixel within the spatially valid set: $R_{val} = \left\{(m, n) \in \mathbb{Z}^2 | t < m \leq M - t, t < n \leq N - t\right\}$. The shift operation in direction α is denoted as S^α, and cropping operation is denoted as C. They are linear operations, therefore gradients can be easily derived for network training.

Distance Layer. It computes feature distance of each shifted feature $S^\alpha(F(X_i))$ of atlas X_i and the target's feature $F(T)$ at each pixel p within R_{val}:

$$D_p^\alpha(T, X_i) = \left\|[C(S^\alpha(F(X_i)))]_p - [C(F(T))]_p\right\|^2. \tag{4}$$

where $[\cdot]_p$ denotes the value, *i.e.*, feature vector, at a pixel p.

Weight Layer. It maps the feature distances to fusion weights using a softmax operation. The fusion weight of pixel q ($q = p - \alpha, \alpha \in R_{nl}$) in atlas X_i for predicting the label of pixel p in target T can be written as

$$w_{i,p,q} = w_p^{\alpha}(X_i) = \frac{e^{-D_p^{\alpha}(T,X_i)}}{\sum_j \sum_{\alpha \in R_{nl}} e^{-D_p^{\alpha}(T,X_j)}} \cdot \tag{5}$$

Voting Layer. It estimates the label of the target image T at pixel p as:

$$\hat{L}_p(T) = \sum_i \sum_{\alpha \in R_{nl}} w_p^{\alpha}(X_i)[C(S^{\alpha}(L(X_i)))]_p \ , p \in R_{val}. \tag{6}$$

Summary: The NL-PLF subnet successively processes atlas and target features/labels by shift, distance, and weight layers to output voting weights, which are further utilized by voting layer to estimate the target label. This subnet implements Eqn.(1) using the above simple layers. For each of them, the gradient of output *w.r.t.* input can be easily derived for efficient network training.

2.3 Network Training

We learn the network parameter Θ by minimizing the loss in Eqn.(3) *w.r.t.* Θ using back-propagation. Given a training set of atlases, each atlas is selected as a target image in turn, and the remaining atlases are taken as the training atlases. If the target image is X_i with ground-truth label $L(X_i)$, the remaining warped atlases are denoted by $\mathcal{A}_i = \{X_j, L(X_j) | j = 1, 2, ..., K, j \neq i\}$. Each triplet of $(\mathcal{A}_i, X_i, L(X_i))$ is called a training data. We use stochastic gradient descent in training, and each training data is taken as a batch. In each batch, we sampled K_0 ($K_0 = 5$) warped atlases as the atlas set, according to a distribution proportional to warped atlas image's normalized mutual information to the target image.

2.4 Atlas Selection in Network Testing

In testing, the learned deep fusion net loads a test sample (a target image and its warped atlases) and outputs the estimated target label. To improve the accuracy, we only pick a few most similar atlases for a target image. Because of the well-trained feature extraction subnet in deep fusion net, we define the distance between a target image and its warped atlas image by:

$$d_F(T, X_i) = \left\| F(T) - F(X_i) \right\|^2, \tag{7}$$

where $F(\cdot)$ is the extracted feature using feature extraction subnet. We take the top-k atlases with least distances as the selected atlases for a target image, as shown in Fig. 3(a). Then the target image and the selected atlases are fed into the learned deep fusion net to produce the estimated target label.

3 Results

3.1 Experimental Setting

Data Set. We apply deep fusion net to the cardiac MR images from MICCAI 2013 SATA Segmentation Challenge for left ventricular segmentation. These subjects are provided by Cardiac Atlas Project, and each subject contains all short-axis cardiac MR images throughout the cardiac cycle with an approximate dimension of $192 \times 192 \times 16 \times 30$.

Image registration. The subject images are with complex backgrounds, we manually crop the ventricular from backgrounds using a bounding box on each subject determined by two corner points, then perform automatic registration and segmentation within the bounding-boxes. To register an atlas to a target image, we perform 3D affine registration for all slices, followed by 2D affine registration and 2D B-spline registration on each slice using ITK [2].

Network structure. In feature extraction subnet, we use four successive repetitions of convolutional layer and ReLU nonlinearity, and finally followed by a normalization layer. Theses convolutional layers respectively have 64 filters in size of $5 \times 5 \times 1$, 64 filters in size of $5 \times 5 \times 64$, 128 filters in size of $5 \times 5 \times 64$, and 128 filters in size of $5 \times 5 \times 128$. Therefore, the extracted feature for each pixel is a 128-D vector. In NL-PLF subnet, we use 7×7 search window. This setting enables accurate performance while taking moderate GPU memory.

We evaluate deep fusion net on end diastolic (ED) frame of 83 training subjects using 5-fold cross-validation, and take the average Dice metric over the validation sets in five times as final accuracy. In each fold, a subset is taken as validation set and the other four subsets are taken for learning deep fusion net.

3.2 Experimental Results

Figure 3(a) shows the top-5 atlases selected by normalized mutual information (NMI) and deep features extracted by feature extraction subnet respectively, while Fig. 3(b) illustrates the mean Dice metrics of searched atlases *w.r.t.* their ranking indexes. The two curves clearly show that our atlas selection method is effective in searching similar atlases for target image compared to NMI.

We compare deep fusion net with majority voting (MV) and state-of-the-art multi-atlas methods for left ventricular segmentation: patch-based label fusion (PB) [2], multi-atlas patch match (MAPM) [10] and SVM with augmented feature (SVMAF) [1]. For fair comparison, all the target images and warped atlases are same for different methods using registration in section 3.1. The results of MV, PB, SVMAF and MAPM are produced by the published codes [3]. Our binary segmentation masks are simply generated by thresholding the fused label

[2] `http://www.itk.org/`

[3] MV, PB, SVMAF: `http://wp.doc.ic.ac.uk/wbai/software/`
MAPM: `https://github.com/BioMedIA/IRTK`

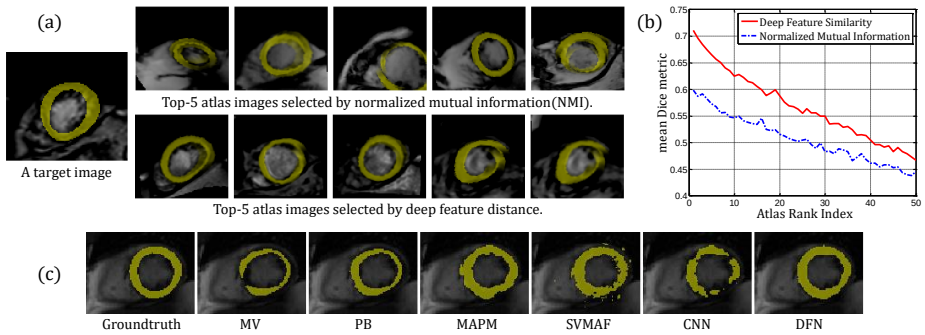


Fig. 3. (a) comparison of atlas selection using NMI and deep features. (b) the mean accuracies of atlases *w.r.t.* their ranking indexes. (c) comparison of segmentation results using different methods.

map using a fixed threshold of 100. As shown in Table 1, our method produces significantly higher accuracy (0.816), and our net using NMI in atlas selection (DFN_NMI) achieves lower result of 0.803. In addition, Fig. 3(c) shows some comparative results. Table 2 shows our results using different number of atlases for each target image, and our method is robust to the number of selected atlases.

Method	MV	PB [2]	MAPM [10]	SVMAF [1]	CNN	DFN_NMI	DFN
Accuracy	0.653	0.683	0.754	0.726	0.681	0.803	**0.816**

Table 1. The mean Dice metrics of different methods.

In Table 1, we also compare deep fusion net to a traditional convolutional neural network (CNN), which has the same net structure as our feature extraction subnet followed by a soft-max layer for classifying each pixel. CNN directly learns a mapping from MR image to label without registration, achieving 0.681 compared to ours (0.816). This shows the advantage of deep fusion net that relies on atlas to target image registration for providing global matching constraint.

Compared to registration method using five landmarks in [1], our registration method only relies on a rough bounding-box, and therefore produces less accurately registered atlases. Notably, our method works significantly better than others, benefiting from the robust atlas selection and effective label fusion enabled by the discriminatively learned deep features.

4 Conclusion

We propose a novel deep fusion net for atlas selection and label fusion in multi-atlas segmentation. Compared to traditional NL-PLF methods, we discriminatively learn optimal deep features for label fusion. Compared to a common CNN for classification, our net relies on the atlas to target image registration. We have

Numb.	1	3	5	7	9	11	13	15	17	19
Accuracy	0.7776	0.8079	0.8141	0.8151	0.8157	0.8161	0.8161	0.8160	0.8158	0.8157

Table 2. The accuracies of DFN using different number of selected atlases.

shown its advantages in Cardiac MR image segmentation. Its success also motivates us to further investigate deep features in registration and segmentation.

Acknowledgments. This work is supported by the NSFC (61472313, 11131006, 11401464) and the 973 program (2013CB329404).

References

1. Bai, W., Shi, W., Ledig, C., Rueckert, D.: Multi-atlas segmentation with augmented features for cardiac mr images. Med. Image Anal. 19(1), 98–109 (2015)
2. Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54(2), 940–954 (2011)
3. Dhungel, N., Carneiro, G., Bradley, A.P.: Deep learning and structured prediction for the segmentation of mass in mammograms. In: Navab, N., Hornegger, J., Wells, M.W., Frangi, F.A. (eds.) MICCAI 2015, Part I. LNCS, vol. 9349, pp. 605–612. Springer, Heidelberg (2015)
4. Duc, A.K.H., Modat, M., Leung, K.K., Cardoso, M.J., Barnes, J., Kadir, T., Ourselin, S.: Using manifold learning for atlas selection in multi-atlas segmentation. PloS one 8(8), e70059 (2013)
5. Giraud, R., Ta, V.T., Papadakis, N., Manjón, J.V., Collins, D.L., Coupé, P.: An optimized patchmatch for multi-scale and multi-feature label fusion. NeuroImage 124, 770–782 (2016)
6. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: A survey. Med. Image Anal. 24(1), 205–219 (2015)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105. Curran Associates, Inc. (2012)
8. 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, M.W., Frangi, F.A. (eds.) MICCAI 2015, Part I. LNCS, vol. 9349, pp. 556–564 (2015)
9. Sanroma, G., Wu, G., Gao, Y., Shen, D.: Learning to rank atlases for multiple-atlas segmentation. IEEE Trans. Med. Imag. 33(10), 1939–1953 (2014)
10. Shi, W., Caballero, J., Ledig, C., Zhuang, X., Bai, W., Bhatia, K.K., et al.: Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part III. LNCS. vol. 8151, pp. 9–16 (2013)
11. Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J.V., Rueckert, D.: Discriminative dictionary learning for abdominal multi-organ segmentation. Med. Image Anal. 23(1), 92–104 (2015)