AtlasNet: Multi-atlas non-linear deep networks for medical image segmentation

M. Vakalopoulou*[1], G. Chassagnon*[1 3], N. Bus[4], R. Marini[4], E. I. Zacharaki[2], M.-P. Revel[3], N. Paragios[1 4]

[1]CVN, CentraleSupélec, Université Paris-Saclay, France
[2]University of Patras, Greece
[3]Groupe Hospitalier Cochin-Hotel Dieu, Université Paris Descartes, France
[4]TheraPanacea, Paris, France

Abstract. Deep learning methods have gained increasing attention in addressing segmentation problems for medical images analysis despite challenges inherited from the medical domain, such as limited data availability, lack of consistent textural or salient patterns, and high dimensionality of the data. In this paper, we introduce a novel multi-network architecture that exploits domain knowledge to address those challenges. The proposed architecture consists of multiple deep neural networks that are trained after co-aligning multiple anatomies through multi-metric deformable registration. This multi-network architecture can be trained with fewer examples and leads to better performance, robustness and generalization through consensus. Comparable to human accuracy, highly promising results on the challenging task of interstitial lung disease segmentation demonstrate the potential of our approach.

Keywords: Encoder-decoder, pixel-wise classification, deformable registration, interstitial lung disease

1 Introduction

Image segmentation is one of the most well studied problems in medical image analysis [8, 6]. Segmentation seeks to group together voxels corresponding to the same organ, or to the same tissue type (healthy or pathological). Existing literature can be classified into two distinct categories, model-free and model-based methods. Model-based methods assume the manifold of the solution space can be expressed in the form of a prior distribution, with sub-space approaches (e.g. active shapes), probabilistic or graphical models and atlas-based approaches being some representatives in this category. Model-free approaches on the other hand rely purely on the observation space combining image likelihoods with different classification techniques.

The emergence of deep learning as disruptive innovation method in the field of computer vision has impacted significantly the medical imaging community [13].

* Authors with equal contribution.

Numerous architectures have been proposed to address task-specific segmentation problems with the currently most successful technique being the Fully Convolutional Network (FCN) [11]. Additionally, FCNs have been combined with upsampling layers, creating a variety of networks [9, 2], and have been extended to 3D [7], boosting even more the accuracy of semantic segmentation.

The main challenges for deep learning in medical imaging arise from the limited availability of training samples – that is amplified when targeting 3D architectures –, the lack of discriminant visual properties and the three-dimensional nature of observations (high dimensional data). In this paper, we propose a novel multi-network architecture that copes with the above limitations. The central idea is to train multiple redundant networks fusing training samples mapped to various anatomical configurations. These configurations correspond to a representative set of observed anatomies and are used as reference spaces (frequently referred to as atlases). The mapping corresponds to a non-linear transformer. Elastic registration based on a robust, multi-metric, multi-modal graph-based framework is used within the non-linear transformer of the network. Training is performed on the sub-space and back-projected to the original space through a de-transformer that applies an inverse nonlinear mapping. The responses of the redundant networks are then combined to determine the optimal response to the problem.

The proposed framework relates also to the multitask learning (MTL) paradigm, where disparate sources of experimental data across multiple targets are combined in order to increase predictive power. The idea behind this paradigm is that by sharing representations between related tasks, we can improve generalization. Even though an inductive bias is plausible in such paradigms, the implicit data augmentation helps reducing the effect of the data-dependent noise. The idea of MTL for image segmentation has been incorporated before, such as in deep networks [10] where soft or hard parameter sharing of hidden layers is performed, or in multi-atlas segmentation [6], where multiple pre-segmented atlases are utilized in order to better capture anatomical variation. As in most ensemble methods, the concept is that the combination of solutions by probabilistic inference procedures can offer superior segmentation accuracy.

AtlasNet differs from previous methods in respect to both scope and applicability. In (single or multi) atlas segmentation, the aim is to map a pre-segmented region of interest from a reference image to the test image, therefore applicability is limited to normal structures (e.g. organs of the body or healthy tissue) that exist in both images. Exploitability is even more reduced in the case of multi-atlas segmentation due to the rareness of multiple atlases. The proposed strategy on the contrary is suitable also for semantic labelling of voxels (as part of healthy or pathological tissue) without the requirement of spatial correspondence between those voxels in atlas and test image.

AtlasNet uses multiple forward non-linear transformers that map all training images to common subspaces to reduce biological variability and a backward de-transformer to relax the effect of possible artificial local deformations. In fact, due to the ill-posedness of inter-subject image registration, regularization

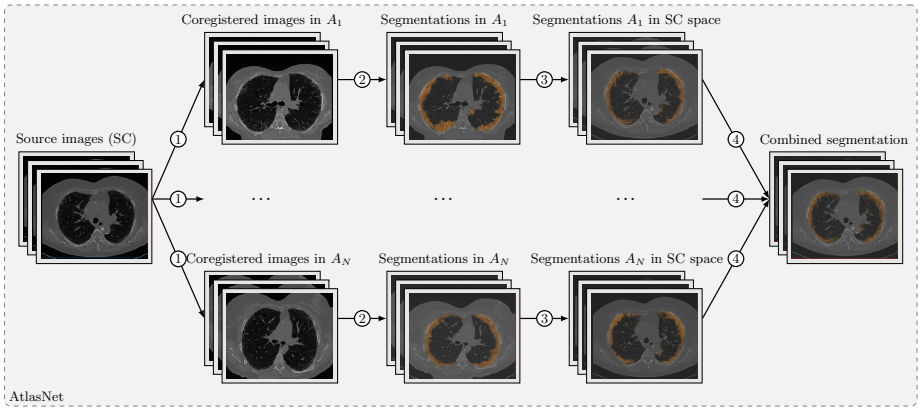


Fig. 1: The proposed AtlasNet framework. A_i indicates atlas i.

constraints are applied to derive smooth solutions and maintain topological relationships among anatomical structures. Consequently, image registration does not always produce a perfectly conforming diffeomorphism due to the nonexistence of a single atlas that matches all anatomies. The use of multiple spaces comes to reduce the atlas selection bias, while the backward transformation aims to balance the effect of possible alterations in local image texture due to the non-linearity in the transformation. Highly promising results comparable to human accuracy on the challenging task of interstitial lung disease (ILD) segmentation demonstrate the potential benefits of our approach. Furthermore, the obtained performance outreached redundant conventional networks.

Finally, the proposed approach addresses most of the limitations of existing neural network approaches. First, it requires fairly small number of training examples due to the reduced diversity of observations once mapped to a common anatomy. Second, it performs data augmentation in a natural manner thanks to the elastic mapping between observations and representative anatomies. Third, it inherits robustness, stability and better generalization properties for two reasons: the limited complexity of observations after mapping, and the "anatomically" consistent redundancy of the networks.

2 Methodology

The method consists of two main parts, a transformer and a de-transformer part. The former maps a sample S to N different atlases $A_i, i \in \{1, \ldots, N\}$, constructs their warped versions, and trains N different networks, while the latter projects back the N predictions to the initial space. These projections are then combined to obtain the final segmentation. The transformer part consists of a non-linear deformable operator (transformer T_i) and a segmentation network C_i while the de-transformer part uses the inverse deformable operator (de-transformer T_i^{-1}) to map everything back to the initial space of a sample S. The framework is

flexible, enables any suitable transformation operator (with an existing inverse) to be coupled with a classifier. The inference framework is illustrated in Fig. 1.

2.1 Multimetric Deformable Operator

The multimetric deformable operator, responsible for mapping samples to different anatomies (atlases) therefore reducing variance and producing anatomically meaningful results, is an elastic image registration method that follows a context-driven metric aggregation approach [4] which aims to find the optimal combination of different similarity metrics. The elastic operator is implemented using a deformable mapping from a source image S to a given atlas A_i. Let us consider that a number of metric functions ρ_j, $j \in \{1, \ldots, k\}$, can be used to compare the deformed source image and the target anatomy A_i. The non-linear transformer T corresponds to the operator that optimizes in the domain Ω the following energy:

$$E(\hat{T}; S, A_i) = \iint_{\Omega} \sum_{j=1}^{k} w_j \rho_j(S \circ \hat{T}, A_i) d\Omega + \alpha \iint_{\Omega} \psi(\hat{T}) d\Omega$$

where w_j are linear constraints factorizing the importance of the different metric functions, and $\psi()$ is a penalty function acting on the spatial derivatives of the transformation as regularization to impose smoothness. Such a formalism can be considered either in the continuous setting that requires differentiable functions with respect to the metric functions ρ_j or in a discrete setting. The advantage of a discrete variant is that it can integrate an arbitrary number and nature of metric functions as well as regularizers while offering good guarantees concerning the optimality properties of the obtained objective function. Inspired by the work done in [5] we express the non-linear operator as a discrete optimization problem acting on a quantized version of the deformation space.

We used free form deformations as an interpolation strategy, invariant to intensity image metrics, pyramidal implementation approach for the optimization and belief propagation for the estimation of the optimal displacement field in the discrete setting. Details on the implementation can be found in [5].

2.2 Segmentation Networks

The segmentation networks C_i operate on the mapped image, $T_i(S)$, to produce a segmentation map and can be the same or different depending on the task and the application and are completely independent of the exact classifier. After defining the optimal deformations $T_i, i = 1..N$, between the source image and the different atlases in the transformer part, AtlasNet uses the inverse transformations to project back to the initial space of the source image S the predicted segmentation maps: $S_i^{seg} = T_i^{-1}(C_i(T_i(S)))$

In this work, motivated by the state-of-the-art performance of F-CNNs in several problems we adapted them for dense labeling. We use the SegNet deep

learning network [2] which performs pixelwise classification and is composed of an encoder and a decoder architecture and follows the example of U-net [9]. It consists of 13 layer groups, similar to the ones of the VGG16 network.The entire architecture consists of repetitive blocks of convolutional, batch normalization, rectified-linear units (ReLU) and indexed max-pooling layers. For more details we refer to the original publication.

Different fusion strategies can be used for the combination of the segmentations. We used the probabilistic output of the classifiers (before hard decision) and fused the output of the different networks based on majority voting.

3 Implementation Details

For the registration, we used the same parameters for all images and all atlases. Three different similarity metrics have been used, namely, mutual information, normalized cross correlation and discrete wavelet metric. For the mutual information 16 bins were used, in the range of -900 to 100.

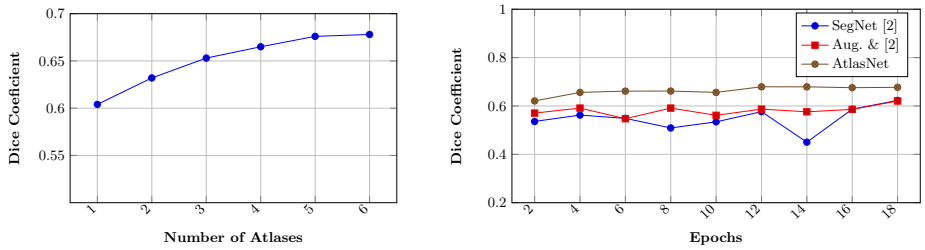
We used the same parameters for training all SegNet networks (initial learning rate = 0.01, decrease of learning rate = $2.5{\cdot}10^{-3}$ every 10 epochs, momentum=0.9, weight decay =$5{\cdot}10^{-4}$). The training of a single network required around 16 hours on a GeForce GTX 1080 GPUs, while the prediction for a single (volumetric) dataset lasted only a few seconds.For data augmentation we performed only random rotations (between -10 and 10 degrees) and translations (between 0 and 20 pixels per axis) avoiding local deformations since the anatomy should not artificially change. Moreover, for training, we performed median frequency balancing [2] to balance the data, as the samples with disease are considerably fewer than the rest of the samples.

4 Experimental Results & Dataset

We used as case study to evaluate our method the ILD segmentation in CT images because it is a challenging problem; boundaries are difficult to detect and delineation suffers from poor-to-moderate interobserver agreement [3]. Moreover, although several visual scoring systems have been proposed to quantify the disease, they only allow basic quantification of ILD severity. The dataset includes 17

Method	Sensitivity	Precision	Hausdorff Dist.	Average Dist.	Dice
SegNet [2]	0.348	**0.623**	4.984	1.891	0.533
Augmentation & SegNet [2]	0.534	0.567	4.077	1.309	0.619
Inter-observer	**0.693**	0.522	4.005	1.317	0.662
AtlasNet	0.682	0.545	**3.981**	**1.274**	**0.677**

Table 1: Different evaluation metrics for the testing dataset for scleroderma disease.



(a) Dice coefficient for different number of atlases.

(b) Dice coefficient evolution for the different methods in the test dataset.

Fig. 2: Quantitative evaluation of the dice coefficient for the proposed method and different number of atlases.

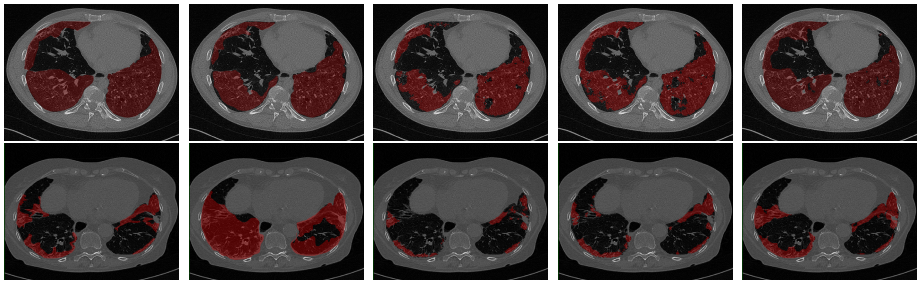
(volumetric) CT images consisting of 6000 slices in total, each being of 512×512 dimension, and annotations of lung and disease. The ILD annotation was performed by a medical expert by tracing the disease boundaries in axial view over all slices and used for training the classification model. Assessment of the method was performed on images from 29 additional patients being fully annotated only on selected CT slices ($n = 20$) by three different observers. Note that the data were multi-vendor (GE & Siemens) and correspond to the same moment of the respiratory cycle.

For all experiments we used six different atlases and registered both training and testing images to them. The choice of atlases was made by a radiologist towards integrating important variability of the considered anatomies. Our experimental evaluation has two objectives: The first one is to show that AtlasNet provides more robust and accurate solutions compared to conventional networks and the second is to examine whether the proposed methodology can truly be trained with fewer examples while leading to good performance. We used five metrics, namely sensitivity, precision, Hausdorff, average contour distance and dice coefficient (over the number of epochs), to evaluate the performance of the proposed method.

On the number of atlases: Fig. 2a presents the behavior of our method using different number of atlases. It can be observed that the dice initially increases and tends to stabilize for more than 5 templates.

Note that, even with the use of only one atlas the deformable operator of AtlasNet helps to increase the dice coefficient (from 0.533 to 0.604), as indicated by Fig. 2b and achieves the highest values of dice compared to conventional networks and usual data augmentation techniques.

On the number of training samples: To evaluate the performance of our architecture with less samples we used a reduced number of samples (30%, 50% and 70% respectively) for the same number of epochs (18) and compare the performance with the one in [2]. The obtained mean dice coefficient values in [2] were 0.434, 0.462, 0.487, while for AtlasNet were 0.613, 0.646 and 0.672 respec-



(a) **Expert 1** (b) **Expert 2** (c) SegNet [2] (d) Augm. & [2] (e) AtlasNet

Fig. 3: Interstitial lung disease segmentation (depicted with red color) on two testing subjects using the different employed strategies.

tively, indicating the robustness of AtlasNet with a significantly lower number of samples. In simple words, the proposed architecture produces better or similar results with 30% of the samples compared to the state-of-the-art architecture [2] with and without data augmentation.

Comparison with the state-of-art: Although results on different datasets are not directly comparable, we compare our method with works related to ILD segmentation. Anthimopoulos *et al.* [1] classified CT image patches with ILD patterns using a CNN and obtained 0.856 accuracy for 6 disease classes. By extracting patches on our data (where different patterns are annotated as a single class) in the same way as in [1] we obtained 0.916 mean accuracy. In [12] a patch-based CNN was augmented with a deep encoder-decoder to exploit partial annotations. By applying AtlasNet on the same dataset as in [12], we increased the mean dice from 0.671 to 0.725.

Moreover we compared AtlasNet in respect to disease segmentation with standard frameworks (without registration and with or without data augmentation) for the same number of epochs (18) and illustrate results in Table 1 and Fig. 2b. For equal comparison, we assessed accuracies using the same classification strategy [2] trained on the initial CT slices, and after performing data augmentation as described earlier. The proposed method reports the best accuracy with respect to Hausdorff distance, average contour distance and dice, indicating that the disease segmentation is much more accurate than by the conventional frameworks with or without data augmentation. This can be inferred also from Fig. 3 where axial slices of two different subjects are depicted. It is clear that the proposed approach segments accurately the boundaries of the disease.

For a more complete evaluation, we compare AtlasNet also with inter-observer agreement using the annotations of three different medical experts. In particular, the annotations of one observer have been used as ground truth to evaluate the rest. From Table 1 and Fig. 3, it can be observed that AtlasNet demonstrates more robust performance than manual segmentation. Finally, it is worth mentioning that even if the network operates on 2D slices, without accounting for

out-of-slice connections, the fusion of the different atlases' predictions makes the final segmentation smooth across all three axes.

Concerning the computational resources, we use a single segmentation network [2] for each of the N atlases, therefore the time and memory usage for one atlas is that of the CNN, while we also showed that a small N (such as 6) suffices. For segmentation of one volumetric CT on a single GPU the total testing time (using 6 atlases) is 3-4min, which includes the registration step. The registration cost is negligible since a graph-based GPU algorithm is used taking 3-5sec per subject. This cost drops linearly with the number and computing power of GPUs. Thus we believe that the additional complexity of AtlasNet is fully justified, since it improves performance by more than 20% and also maintains it stable with only 30% of the training data compared to conventional single networks.

5 Conclusion

In this paper, we present a novel multi-network architecture for (healthy or pathological) tissue or organ segmentation that maximizes consistency by exploiting diversity. Evaluation of the method on interstitial lung disease segmentation highlighted its advantages over previous competing approaches as well as inter-observer agreement. The investigation of techniques for soft parameter sharing of hidden layers, and information transfer between the different networks and atlases is our direction for future work. Finally, the extension to multi-organ segmentation including multiple classes loss functions is one of the potential directions of our method.

6 Acknowledgements

This work has been partially supported by the european project ERC-PoC 737604 TheraPanacea.

References

1. Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., Mougiakakou, S.: Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. IEEE Trans Med Imaging 35(5) (2016)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE PAMI (2017)
3. Camiciottoli, G., et al.: Lung ct densitometry in systemic sclerosis: Correlation with lung function, exercise testing, and quality of life. Chest 131(3) (2007)
4. Ferrante, E., Dokania, P.K., Marini, R., Paragios, N.: Deformable Registration Through Learning of Context-Specific Metric Aggregation, pp. 256–265. Springer International Publishing, Cham (2017)
5. Glocker, B., Komodakis, N., Tziritas, G., Navab, N., Paragios, N.: Dense image registration through mrfs and efficient linear programming. Medical Image Analysis 12(6), 731 – 741 (2008)

6. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: A survey. Medical Image Analysis 24(1), 205 – 219 (2015)
7. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on. pp. 565–571. IEEE (2016)
8. Paragios, N., Ferrante, E., Glocker, B., Komodakis, N., Parisot, S., Zacharaki, E.I.: (hyper)-graphical models in biomedical image analysis. Medical Image Analysis 33 (2016)
9. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer (2015)
10. Ruder, S.: An overview of multi-task learning in deep neural networks. CoRR abs/1706.05098 (2017)
11. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(4), 640–651 (April 2017)
12. Vakalopoulou, M., Chassagnon, G., Paragios, N., Revel, M., Zacharaki, E.: Deep patch-based priors under a fully convolutional encoder-decoder architecture for interstitial lung disease segmentation. In: 2018 IEEE International Symposium on Biomedical Imaging (ISBI) (2018)
13. Zhou, S., Greenspan, H., Shen, D.: Deep Learning for Medical Image Analysis. Academic Press (2017)