Probabilistic Atlases to Enforce Topological Constraints

Udaranga Wickramasinghe¹, Graham Knott², and Pascal Fua¹

¹ Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

udaranga.wickramasinghe@epfl.ch

² BioEM Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

Abstract. Probabilistic atlases (PAs) have long been used in standard segmentation approaches and, more recently, in conjunction with Convolutional Neural Networks (CNNs). However, their use has been restricted to relatively standardized structures such as the brain or heart which have limited or predictable range of deformations. Here we propose an encoding-decoding CNN architecture that can exploit rough atlases that encode only the topology of the target structures that can appear in any pose and have arbitrarily complex shapes to improve the segmentation results. It relies on the output of the encoder to compute both the pose parameters used to deform the atlas and the segmentation mask itself, which makes it effective and end-to-end trainable.

1 Introduction

Probabilistic atlases (PAs) are widely used for multi-atlas based segmentation [8]. With the advent of deep learning, there has been a push to incorporate them with Convolutional Neural Networks (CNNs) as well [1,7,14,15]. The published techniques relying on deep learning share a number of features: They work best for structures featuring relatively small variations in shape and position; the atlases are often created by fusing multiple manually annotated images; the atlases must also be pre-registered to the target images to align them with the structures of interest.

Thus, the PAs have been mostly used to segment structures such as the brain or heart for which the above requirements can be met. In this paper, we propose an approach to design and handle PAs that makes them usable in complex situations where the shape and position can vary dramatically, such as those shown in Fig. 1. Our PAs are coarse and only encode the relative position and topology of the structures we expect to find. To register them, we rely on affine transforms whose parameters are estimated at the same time as the segmentations themselves using the end-to-end trainable encoder-decoder architecture depicted by Fig. 2, which we will refer to as **PA-Net**. The affine transforms are used to warp the atlases and feed the features of the warped atlases to the decoder. This differs significantly from earlier approaches [14,15] that rely on pre-registered atlases.

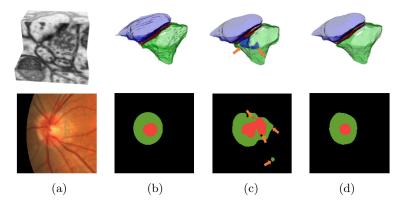


Fig. 1. Eliminating topological mistakes. (a) A small FIBSEM image stack featuring a synaptic junction and a retinal fundus image. (b) Ground-truth segmentations. The pre-synaptic, synaptic-cleft, and post-synaptic regions shown in green, red, and blue respectively. Similarly, the optic disk and cup shown in green and red, respectively. (c) U-Net segmentations with orange arrows pointing to topological mistakes. (d) Our error-free result (best viewed in color).

We validate our approach on segmentation of synaptic junctions in Electron Microscopy (EM) images and optic nerve head segmentation in retinal fundus images. A synaptic junction has an arbitrary shape but always features a synaptic cleft sandwiched between a pre-synaptic bouton belonging to the axon of one neuron and a post-synaptic dendritic spine belonging to another. In this case, the shape complexity and unpredictability precludes the use of standard PAs. Optic nerve head consists of an optic disk and optic cup. Even though their shape does not vary significantly, there are significant variations in the size of the optic cup and the position of the optic nerve head.

We will show that **PA-Net** eliminates most of the topological mistakes its unaided V-Net [11] or U-Net [13] backbone makes on the EM and retinal images, while being generic and potentially applicable to other segmentation tasks. Our contribution is therefore a novel approach that makes it possible to use probabilistic atlases even in situations where shape variability is too great for existing approaches and methods to integrate the atlas registration process directly into the network. The code is publicly available³.

2 Related Work

PAs are commonly used in standard atlas-based segmentation approaches. In their piplines, CNNs are indirectly used to perform segmentation. These include using a CNN to infer deformation maps to register target images to atlas images [2] and using it to implicitly infer the segmentation masks[4,6]; using CNNs to

³ https://github.com/cvlab-epfl/PA-net.git

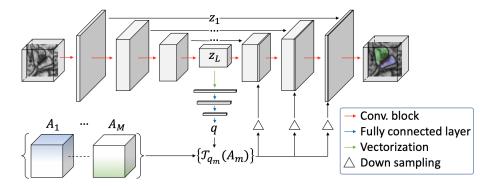


Fig. 2. PA-Net architecture. The network takes as input an image and a list of PAs. Given the decoder output, the fully connected layers estimate the parameters of the affine transformation that registers the PAs to the input. The registered PAs are then concatenated with the features computed by the convolutional decoder at each scale.

improve the selection of best atlases from a library [10]. As this is not the focus of this paper, we will not discuss them further and will concentrate on methods that directly use the PAs to assist a CNN that perform segmentation.

A PA can provide localization priors to help a network find the target of interest. Such PAs are often referred to as seed layers. They come in two flavors, Gaussian priors [3,16] or binary seed layers [9]. Another popular way to focus the attention of the network is to provide an attention mask [5], which then serves the same purpose as a PA. However, the attention masks are derived from the input data itself and does not act as an external knowledge source.

PAs built by fusing multiple manually annotated images can be used to introduce even more prior knowledge about shape and topology of structures. In [14,15], this is done to improve segmentations of human brain by having the CNN take pre-registered PAs as input. In our work, we extend the idea by demonstrating that PAs can be used to introduce topological knowledge even when the structure of interest exhibits large shape variations and without pre-registration that the whole system is end-to-end trainable.

3 Approach

3.1 Network Architecture

The architecture of the proposed **PA-Net** is depicted in Fig. 2. Its backbone is an encoder-decoder architecture similar to the one used in popular networks such as U-Net [13] or V-Net [11]. These networks are designed to learn to approximate the distribution $p(\mathbf{Y}|\mathbf{X})$, where \mathbf{X} is the input image and \mathbf{Y} is the output segmentation. We extend the idea by learning the joint distribution $p(\mathbf{Y}, T_{\mathbf{q}}(\mathcal{A})|\mathbf{X}, \mathcal{A})$. The function $\mathcal{T}_{\mathbf{q}}(\cdot)$ applies affine transformations to the set of PAs \mathcal{A} given the

pose vectors q. We factorize this joint distribution as

$$p(\mathbf{Y}, T_{\mathbf{q}}(\mathcal{A})|\mathbf{X}, \mathcal{A}) = p(\mathbf{Y}|\mathbf{X}, T_{\mathbf{q}}(\mathcal{A}))p(\mathbf{q}|\mathbf{X}, \mathcal{A})$$
(1)

PA-Net models it using a primary stream, the encoder-decoder backbone at the top of Fig. 2, and a secondary stream, the fully connected layers at the bottom of Fig. 2. The secondary stream $\mathcal{F}(\cdot, \theta_f)$ originates from the latent vector produced by the final stage of the encoder $\mathcal{E}(\cdot, \theta_e)$, estimates the affine parameter vector \mathbf{q} , uses it to warp the atlases, and feed them to the various layers of the decoder $\mathcal{D}(\cdot; \theta_d)$ after rescaling them to the resolution of feature vectors at each stage. Here θ_e, θ_d and θ_f are the learned weights that control the behavior of \mathcal{E}, \mathcal{D} and \mathcal{F} . Given the encoder output, \mathcal{F} and \mathcal{D} model $p(\mathbf{Y}|\mathbf{X}, T_{\mathbf{q}}(A))$ and $p(\mathbf{q}|\mathbf{X}, A)$, respectively. This guarantees that these two distributions are learned from the same features. Training the encoder to produce features that are effective for both tasks yield improved performance.

3.2 Atlas Design

Fig 3. depicts the topological atlases we use to segment synaptic junctions and optic disks and cups. In the first case, they are cubes and in the second, ordinary 2D arrays. They encode the probability of any pixel of voxel to belong to one of the possible classes—pre-synaptic, cleft, or post-synaptic for synapses and optic disk or cup for retinas—given its pose.



Fig. 3. Probabilistic atlases. (a) For synaptic junction segmentation, they are cubes and we show a single face. (b) For optic optic disk and cup segmentation, they are regular 2D arrays. The color hues denote the same areas as those in Fig. 1. The color saturation is proportional to the probability at a given location (best viewed in color).

3.3 Atlas Registration

Parameters necessary for PA registration is computed by estimating the pose vector q. When working with image cubes such as the one shown at the top of Fig. 1, we take the pose vector to be $\mathbf{q} = [t_x, t_y, t_z, s_x, s_y, s_z, r_x, r_y, r_z]$, where $t_{\{x,y,z\}}$ represent translations and $s_{\{x,y,z\}}$ scalings in the x,y, and z directions, respectively. We represent the orientation using the angle-axis vector $[r_x, r_y, r_z]$, whose direction is the axis of rotation. When working with 2D images, we drop t_z , s_z , and r_z from \mathbf{q} .

	Jaccard index (%)				
	Pre-synaptic	Synaptic junction	Post-synaptic	Mean	TER
V-Net	63.0	56.0	80.6	66.5	5/15
PA-VNet	59.5	57.8	83.1	66.8	1/15
U-Net	73.6	59.7	79.1	70.7	7/15
Naive PA-UNet	73.4	58.0	78.4	69.9	7/15
PA-UNet	72.6	62.8	87.0	74.1	3/15

Table 1. Comparative results for synaptic junction segmentation.

3.4 Loss Functions

Following standard practice, we formulate the loss terms as the negative log likelihood of the joint distribution $p(\mathbf{Y}, T_{\mathbf{q}}(A)|\mathbf{X}, \mathcal{A})$ of Eq. 1. This yields the composite loss

$$\mathcal{L} = \mathcal{L}_{seq} + \mathcal{L}_{pose} . \tag{2}$$

 \mathcal{L}_{seg} is the standard cross entropy loss that evaluates segmentation performance and take \mathcal{L}_{pose} is the least square error in estimating the pose \mathbf{q} .

4 Results and Discussion

4.1 Datasets

Synaptic Junction Dataset: It is a $500 \times 500 \times 200$ FIB-SEM image stack of a mouse cortex. We used 50 xy slices for training, 50 for validation, and 100 for testing. From each set, we cropped $96 \times 96 \times 96$ image volumes containing a synaptic junction such as the one shown in the top row of Fig. 1(a) and they are zero-padded as necessary. This gave us 13, 10 and 15 volumes for training, validation and testing, respectively. The synapse is not necessarily perfectly centered and the task is to segment pre-synaptic region, post-synaptic region, and synaptic cleft.

Retinal Fundus Image Dataset: It comprises $400\ 2124 \times 2056$ retinal fundus images acquired using a Zeiss Visucam 500 camera [12], which we resize and pad to be 512×512 . We use 100 for training, 100 for validation, and 200 for testing.

4.2 Quantitative and Qualitative Results

We evaluate segmentation performance in terms of the standard Jaccard Index and of an additional metric we dub the *Topological Error Ratio* (TER). We define TER as the ratio of segmentations containing topological errors to the total number of test images. A segmentation is considered to contain a topological error if it violates the expected topology of the target structure, that is, it features semantic labels appearing where they should not given the topological

	Jaccar				
	Optic disk			TER	
V-Net	78.4	72.5	75.5	7/200	
PA-VNet	79.0	73.0	76.0	3/200	
U-Net	81.3	73.8	77.6	18/200	
Naive PA-UNet	80.8	74.1	77.3	16/200	
PA-UNet	81.6	74.6	78.1	5/200	

Table 2. Comparative results for fetinal fundus image segmentation.

constraints. We compute this value by finding the connected components of the final segmentation and then finding instances that violate the topology.

We tested two versions of **PA-Net**, one based on the U-Net [13] architecture and the other based on the more recent V-Net [11]. We compare the results against those of the standard U-Net and V-Net. We report our comparative results on our two datasets in Tables 1 and 2. Figs. 4 depict them qualitatively (see supplementary materials for further results). Using the atlas consistently improves over baseline performance. The gains are most visible in TER terms because our **PA-Nets** truly come into their own when the standard U-Net and V-Net fail, which is only a fraction of the time, as would be expected of architectures that are as popular as they are. Even when it fails to eliminate topological errors, it reduces the size of the errors as shown in the second row of Fig. 4). The one exception is the pre-synaptic region, for which the Jaccard numbers decrease. A closer inspection of the results show that the atlas sometimes encourages the segmentation to leak from weak boundary regions in few instances. We plan to address this issue by introducing an additional stream that specifically focus on predicting object boundaries and combining its results with **PA-Net**.

To demonstrate the importance of using the same features to compute the segmentation and the pose parameters, we implemented a naive version of our approach that uses two separate streams to compute the pose and the segmentation features, which we denote as Naive PA-Unet in Tables 1 and 2. The naive version fails to resolve the topological mistakes in the datasets. In the synaptic junction dataset, this occurs because the naive version has a mean orientation estimation error of 80.6° in contrast to the PA-Net's error of 10.4°. In the retinal fundus image dataset, the naive version has a mean localization error of 26.5 pixels compared to a 3.9 pixel error for PA-Net. As a result, in both instances, network fails to properly register PAs.

5 Conclusion

We have proposed an encoding-decoding architecture that can exploit rough atlases that encode the topology of structures that can appear in any pose and have arbitrarily complex shapes to improve the segmentation results. One of its crucial components is that it relies on the output of the encoder to compute both the pose parameters used to deform the atlas and the segmentation mask itself,

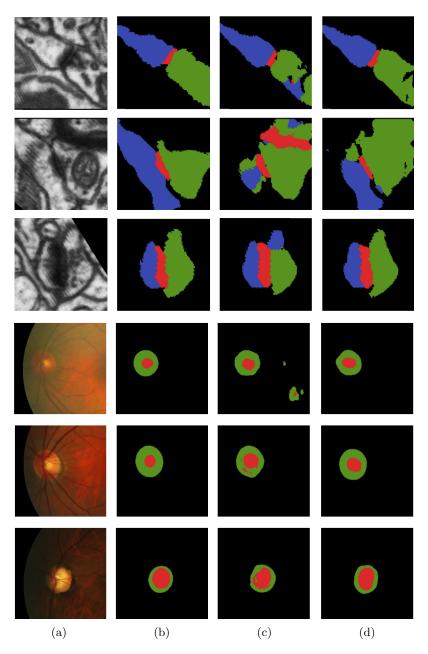


Fig. 4. U-Net vs **PA-Net**. Top three rows depict results from synaptic junction segmentation. The bottom three rows depict results from retinal fundus image segmentation. (a) Input images, (b) Ground-truth, (c) U-Net results (d) **PA-Net** results. The color hues denote the same areas as those in Fig. 1 (best viewed in color).

which makes it effective and end-to-end trainable. As future work, we plan to extend **PA-Net** by introducing an additional stream to produce edge maps and using it to address the minor loss in accaracy that occur when the structures have weak boundaries.

Acknowledgments: This work was supported in part by a Swiss National Science Foundation grant.

References

- Atzeni, A., Jansen, M., Ourselin, S., Iglesias, J.E.: A probabilistic model combining deep learning and multi-atlas segmentation for semi-automated labelling of histology. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS. vol. 11071, pp. 219–227. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_25
- 2. Balakrishnan, G., Zhao, A., Dalca, A., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: CVPR (2018)
- 3. Bertasius, G., Park, H.S., Yu, S.X., Shi, J.: Unsupervised Learning of Important Objects from First-Person Videos. In: ICCV (2017)
- 4. Dong, S., Luo, G., Wang, K., Cao, S., Mercado, A., Shmuilovich, O., Zhang, H., Li, S.: Voxelatlasgan: 3d left ventricle segmentation on echocardiography with atlas guided generation and voxel-to-voxel discrimination. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MIC-CAI 2018. LNCS. vol. volume 11073, pp. 622–629. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_71
- Fei, W., Mengqing, J., Chen, Q., Shuo, Y., Cheng, L., Honggang, Z., Xiaogang, W., Xiaoou, T.: Residual attention network for image classification. In: CVPR (2017)
- 6. Hering, A., Kuckertz, S., Heldmann, S., Heinrich, M.P.: Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking. In: Bildverarbeitung für die Medizin (2019)
- Huo, Y., Xu, Z., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A.: Spatially localized atlas network tiles enables 3d whole brain segmentation from limited data. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS. vol. 11072, pp. 698–705. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_80
- 8. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: A survey. Med Image Anal (2015)
- 9. Januszewski, M., Kornfeld, J., Li, P., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J., Tyka, M., Denk, W., Jain, V.: High-Precision Automated Reconstruction of Neurons with Flood-Filling Networks. Nat. Methods (2018)
- Katouzian, A., Wang, H., Conjeti, S., Tang, H., Dehghan, E., Karargyris, A., Pillai, A., Clarkson, K., Navab, N.: Hashing-based atlas ranking and selection for multiple-atlas segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS. vol. 11073, pp. 543-551. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_62
- 11. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571 (Oct 2016)

- 12. Retinal fundus glaucoma challenge (2018),
 $\mbox{http://refuge.grand-challenge.org},$ last accessed 24 Jun
 2019
- 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). https://doi.org/10.1007/978-3-319-24574-4_28
- Spitzer, H., Amunts, K., Harmeling, S., Dickscheid, T.: Parcellation of visual cortex on high-resolution histological brain sections using convolutional neural networks. In: ISBI. pp. 920–923 (2017)
- Spitzer, H., Kiwitz, K., Amunts, K., Harmeling, S., Dickscheid, T.: Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS. vol. 11072, pp. 663–671. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_76
- Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient Video Object Segmentation via Network Modulation. In: CVPR (2018)