Learning to Segment 3D Linear Structures Using Only 2D Annotations

Mateusz Koziński[*], Agata Mosinska[**], Mathieu Salzmann, and Pascal Fua
{name.surname}@epfl.ch

Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland

Abstract. We propose a loss function for training a Deep Neural Network (DNN) to segment volumetric data, that accommodates ground truth annotations of 2D projections of the training volumes, instead of annotations of the 3D volumes themselves. In consequence, we significantly decrease the amount of annotations needed for a given training set. We apply the proposed loss to train DNNs for segmentation of vascular and neural networks in microscopy images and demonstrate only a marginal accuracy loss associated to the significant reduction of the annotation effort. The lower labor cost of deploying DNNs, brought in by our method, can contribute to a wide adoption of these techniques for analysis of 3D images of linear structures.

1 Introduction

Linear structures such as blood vessels, bronchi and dendritic trees are pervasive in medical imagery. Automatically recovering their topology has therefore become critically important to fully exploit the vast amounts of data that modern imaging devices can now produce. Machine Leaning based techniques have demonstrated their effectiveness for this purpose, but usually require substantial amounts of annotated training data to reach their full potential.

Unfortunately, annotating complex topologies in 3D volumes by means of an inherently 2D computer interface is slow and tedious. The annotator must frequently rotate and move the volume to verify the correct placement of control points and to reveal occluded details. Not only is this inherently slow, but such interactions require continuously re-displaying large amounts of data, which often exceeds the capacity of a workstation, thus introducing further delays.

In this paper, we show that we can train a Deep Net to perform 3D volumetric delineation given *only* 2D annotations in Maximum Intensity Projections (MIP), such as those shown on the right side of Fig. 1. This is a major time-saver because delineating linear structures in 2D images is much easier than in 3D volumes and involves none of the difficulties mentioned above. Furthermore, semi-automated annotation tools work more smoothly on 2D than on 3D data. In short, limiting

[*] the author was supported by the FastProof ERC Proof of Concept Grant
[**] the author was supported by the Swiss National Science Foundation

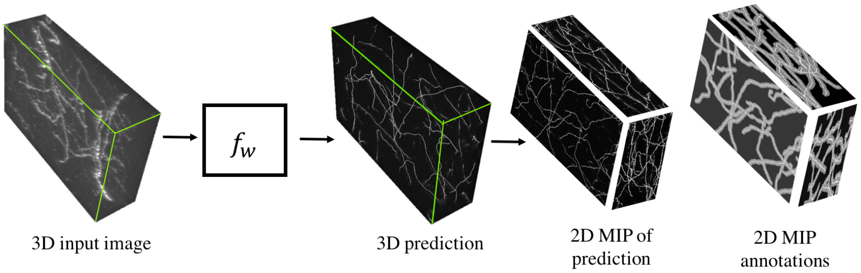

Fig. 1. 3D training using 2D annotations only. We first annotate the 2D Maximum Intensity Projections (MIP) of the training image stacks. Then, we minimize a cross entropy loss between the annotated 2D MIPs and the corresponding projections of the 3D prediction made by the network f_w we are training.

the annotation effort to the projections leads to a considerable labor saving without compromising the performance of the trained network.

More specifically, we introduce a loss function that penalizes discrepancies between the maximum intensity projection of the predictions and the 2D annotations. We show that it yields a network that performs as well as if it had been trained using full 3D annotations. The loss is inspired by *space carving*, a classical approach to reconstructing complex 3D shapes from arbitrarily-positioned cameras [1]. Space carving exploits the fact that visual rays corresponding to background pixels in 2D images cannot cross any foreground voxel when passing through the volume. Conversely, rays emanating from foreground pixels have to cross at least one foreground voxel. In our case, the rays are parallel to the projection axes. The network is trained to minimize the cross-entropy between the 2D annotations and the maximum values along the rays.

Our contribution is therefore a principled approach to reducing the annotators' burden when training a Deep Net by enabling them to trace in 2D instead of 3D, while still capturing the full 3D topology of complex linear structures. We demonstrate this on 3D light microscopy images of neurons and retinal blood vessels and on Magnetic Resonance Angiography (MRA) brain scans.

2 Related Work

Early approaches to delineation of 3D curvilinear structures relied on filters manually designed to respond strongly to tubular segments [2–4]. They do not require to be trained, but their performance degrades when the structures become irregular and the images noisy. This has led to the emergence of machine learning based methods that can cope with such difficulties, given enough annotated data [5–8]. The most recent one of these [8] relies on Deep Learning for neuron tracing by adaptive exploration of 3D light microscopy images.

However, using Machine Learning, and Deep Learning in particular, requires large amounts of annotated training data. Furthermore, annotating 3D stacks is much more labor-intensive than annotating 2D images. Only true experts, whose time is precious, are able to orient themselves and follow complex structures in large volumes [9]. Until now, this problem has been handled by developing better ways to visualize and interact with image stacks [10, 8]. In [11], only a few slices of a volume are annotated and the loss is computed using only them. The technique of [9], like ours, allows the annotator to trace a linear structure in a maximum intensity projection and then attempts to guess the value of the third coordinate using a simple heuristic. While effective when the structures are relatively sparse, this can easily get confused as the scene becomes more cluttered.

The originality of our approach is to introduce a method that relies solely on 2D annotations in Maximum Intensity Projections, yet captures the 3D structure of complex linear structures when the projections are used jointly.

3 Method

3.1 From 3D to 2D Annotations

Let us first consider the problem of training a neural network f_w, parameterized by weights w, to segment linear structures within 3D image stacks, given a training set T of pairs $(\mathbf{x}, \tilde{\mathbf{y}})$, where each 3D image $\mathbf{x}$ is accompanied by the corresponding volumetric ground-truth annotations $\tilde{\mathbf{y}}$. We denote the elements of $\mathbf{x}$ and $\tilde{\mathbf{y}}$ by x_{ijk} and $\tilde{y}_{ijk}$, where i, j, k index the positions of the elements within the volumes. The ground-truth labels take a value in the set $\{1, 0, \varnothing\}$, which indicate the presence of a linear structure in voxel i, j, k if $\tilde{y}_{ijk} = 1$, the absence of a linear structure if $\tilde{y}_{ijk} = 0$, and uncertainty of the annotator if $\tilde{y}_{ijk} = \varnothing$. Delineation can then be cast as a binary segmentation problem by simply ignoring the voxels labeled as $\varnothing$ during training. The network output $\mathbf{y} = f_w(\mathbf{x})$ has the same size as the input and contains probabilities of presence of a linear structure in each voxel. To train the network, we find

$$\arg\min_{w} \sum_{(\mathbf{x},\tilde{\mathbf{y}}) \in T} \sum_{i,j,k} L(f_w(\mathbf{x})_{ijk}, \tilde{y}_{ijk}) \,, \tag{1}$$

where $f_w()_{ijk}$ denotes voxel i, j, k of the prediction, and the loss $L(y, \tilde{y})$ is taken to be the cross entropy $C(y, \tilde{y}) = [\tilde{y} = 1] \log y + [\tilde{y} = 0] \log(1 - y)$, where $[\cdot]$ is the Iverson bracket. As discussed in the introduction, the drawback of this approach is that generating the ground-truth labels $\tilde{\mathbf{y}}$ in sufficient numbers to train a deep network is tedious and expensive when operating on large volumes.

To alleviate this problem, we reformulate the loss function of Eq. 1 so that it can exploit annotated Maximum Intensity Projections (MIPs) of the input volumes. A MIP of volume $\mathbf{x}$ along direction i, which we denote as $\mathbf{x}^{\mathrm{i}}$, is a 2D image with elements $x^{\mathrm{i}}_{jk} = \max_i x_{ijk}$. Annotating MIPs is easy when the structures of interest have high intensity and are clearly visible in the projections. A MIP annotation $\tilde{\mathbf{y}}^{\mathrm{i}}$ comprises elements $\tilde{y}^{\mathrm{i}}_{jk} \in \{1, 0, \varnothing\}$, which can also be thought of as $\tilde{y}^{\mathrm{i}}_{jk} = \max_i \tilde{y}_{ijk}$. MIPs of the volume along the directions j and k, and their annotations, are defined similarly.

Since $\tilde{y}^{\mathrm{i}}_{jk} = 0$ tells us that *all* voxels of the input column jk contain background while $\tilde{y}^{\mathrm{i}}_{jk} = 1$ tells us that at least one voxel in the input column contains

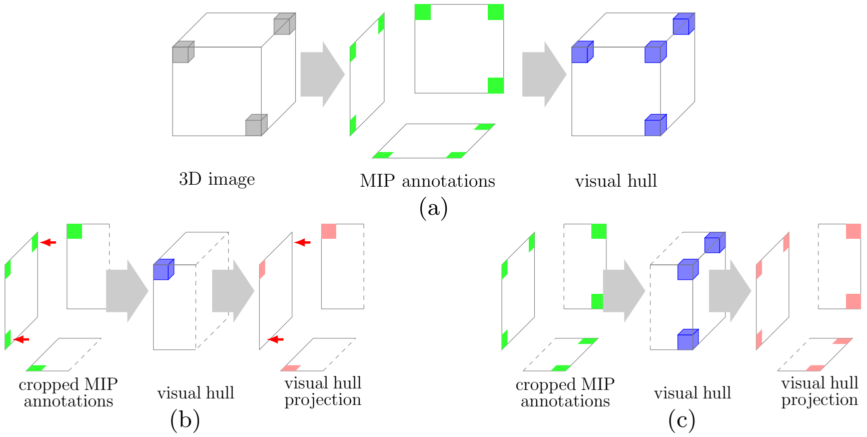

Fig. 2. Handling cropped volumes. *(a)* A 3D volume with three foreground voxels, the annotations of its MIPs in green, and the visual hull computed from these in blue. *(b)* The volume has been cropped so that only the left half remains. The annotations have been cropped to match, leaving a single blue voxel in the visual hull. Reprojecting it into the MIPs lets us eliminate the extraneous annotations, indicated with red arrows. *(c)* However, there are situations such as the one depicted here, where some will survive.

a linear structure, we define the max-projection $f^{\mathrm{i}}_w(\mathbf{x})$ along direction i of the network output as the image with elements $f^{\mathrm{i}}_w(\mathbf{x})_{jk} = \max_i f_w(\mathbf{x})_{ijk}$. We proceed similarly for directions j and k. We then rewrite our training loss as

$$\sum_{(\mathbf{x},\tilde{\mathbf{y}})\in T}\Big(\sum_{jk} L\big(f^{\mathrm{i}}_w(\mathbf{x})_{jk},\tilde{y}^{\mathrm{i}}_{jk}\big)+\sum_{ik} L\big(f^{\mathrm{j}}_w(\mathbf{x})_{ik},\tilde{y}^{\mathrm{j}}_{ik}\big)+\sum_{ij} L\big(f^{\mathrm{k}}_w(\mathbf{x})_{ij},\tilde{y}^{\mathrm{k}}_{ij}\big)\Big)\,. \qquad (2)$$

Note that $f^{\mathrm{i}}_w(\mathbf{x})_{jk}$ upper bounds the probability of presence of a linear structure in column jk. Eq. 2 penalizes large values of this upper bound whenever $\tilde{y}^{\mathrm{i}}_{jk} = 0$, thus mimicking space carving. When $\tilde{y}^{\mathrm{i}}_{jk} = 1$, minimizing the loss increases the largest prediction in the column.

3.2 Visual Hull for Training on Cropped Volumes

Due to memory limitations, the annotated training volumes are typically cropped into sub-volumes and the MIP can be cropped to match. However, the cropped annotations may then contain labels for structures located outside the volume crop, as illustrated by Fig. 2. To reduce the influence of these extraneous annotations, we use another element of the space carving theory, the visual hull **h**. **h** is a volume containing the original one, and constructed from its projections [1].

By construction, an element of the hull $h_{ijk} = 1$ if and only if *all* of its projections are labelled as foreground. In our context, a foreground voxel outside a crop only produces an incorrect annotation in *a single* projection. Therefore, as shown in Fig. 2, we can very often eliminate it by projecting the visual hull back to the 2D annotations and discarding those that fall outside.

3.3 Implementation

In practice, we implemented f_w as a U-Net style network [12]. Specifically, we made the original convolution-ReLU blocks residual, and only used two max-pooling operations instead of the usual four, which resulted in a more compact network that fits in memory even with larger volume crops. In all our experiments, we trained the network for 200K iterations, using the ADAM update scheme [13] with momentum of 0.9, weight decay 10^{-4} and step size 10^{-5}.

4 Experimental Evaluation

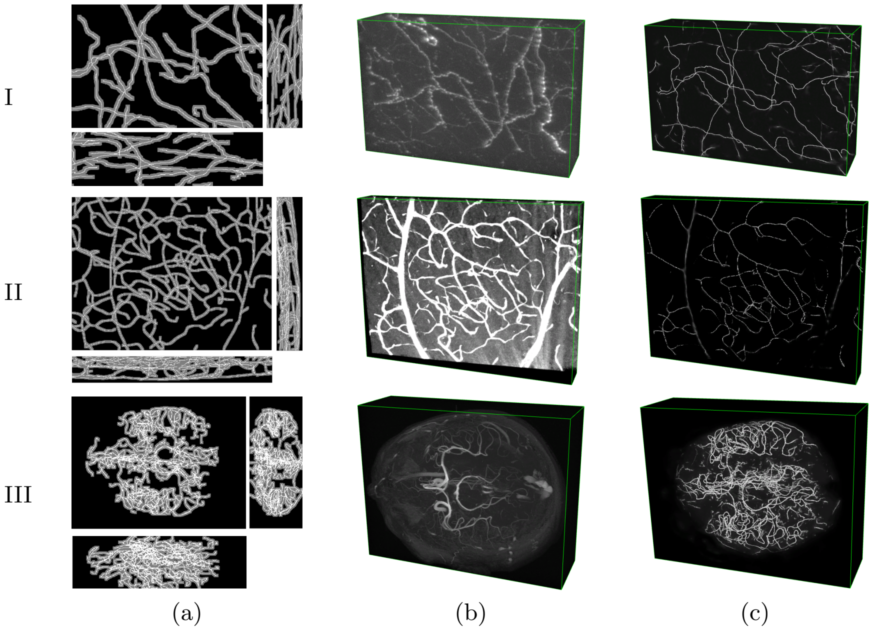

Fig. 3. Results on our three datasets, from top to bottom, axons, retinal blood vessels, and brain vasculature in MRA scans. (a) 2D annotations in 3 MIPs of a training volume. The foreground centerline annotations are marked in white and the regions to be ignored around them in gray. (b) Input test image volume. (c) Output segmentation.

4.1 Data and Annotations

We tested our approach on three data sets that differ in terms of the imaged tissue, the acquisition modality and the image resolution. As a result, there are substantial variations with respect to the density of the structures of interest, their appearance and the amount of clutter originating from extraneous objects.

Axons. The dataset comprises 16 stacks of 2-photon microscopy images of mouse neural tissue, with sizes ranging from $40 \times 200 \times 200$ to $136 \times 322 \times 500$ voxels and a resolution of $0.8 \times 0.26 \times 0.26$ µm. We split the data into a test set of two volumes of size 136 × 233 × 500, and a training set of 14 smaller volumes. The top row of Fig. 3 depicts one of the test volumes.

Retina. The dataset is made of two confocal microscopy image stacks depicting retinal blood vessels, sized 1024 × 1024 × 110, and with a resolution of 0.62 µm. We use one for training and the other, depicted in Fig. 3, for testing. Since most vessels are located within a 50-pixel high XY slice, MIPs in the X and Y directions are very cluttered. Therefore, we split the volume into 16 256 × 256 × 110 subvolumes and annotated their MIPs. In other words, we also traced the vertical faces of the smaller volumes. This only requires annotating 6 additional 1024 × 110 images, which is still fast. The middle row of Fig. 3 describes both our 2D annotations and the results on one of the test sets.

Angiography. This set of MRI brain scans [14], one of which is shown in Fig. 3, is publicly available. It consists of 42 annotated stacks, which we cropped to a size of 416 × 320 × 128 voxels by removing the margins. Their resolution is 0.5 × 0.5 × 0.6 mm. We partitioned the data into 31 training and 11 test volumes. As in the case of the retinal vessels, we decreased the visual clutter by splitting each volume into 4 208 × 160 × 128 subvolumes for which we produced 2D annotations. This requires annotating an additional 416 × 128 image and a 320 × 128 one. The bottom row of Fig. 3 describes both our 2D annotations and our results on one of the test sets.

All the manual annotations are expressed in terms of the 2D and 3D centerlines of the underlying structures. We then use a pixel-width of 11 for the first two datasets and 7 for the third to define the area around the centerline to be ignored when computing the loss, as discussed in Section 3.1, as well as for computation of the visual hulls, as described in Section 3.2.

4.2 User Study

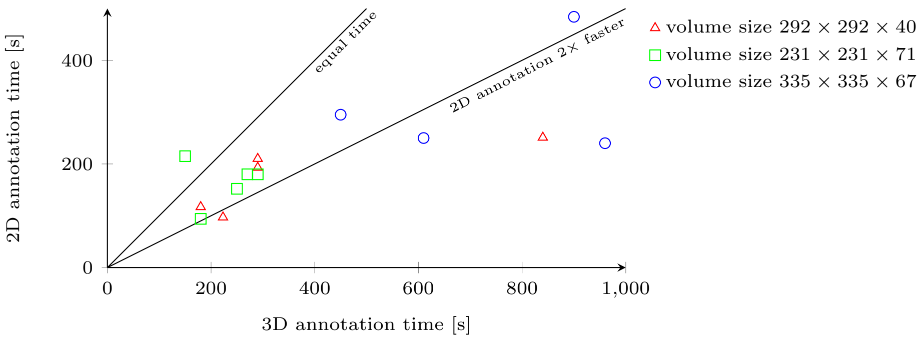

Fig. 4. Annotation times captured during the user study. Each user annotated each of the volumes both in 3D and in 2D. Each point in the plot represents the time consumed by a single user to annotate a single volume, its x-coordinate corresponding to the 3D annotation time, and the y-coordinate representing the time needed to complete the 2D annotation. Different colors denote different volumes.

The usefulness of our approach is predicated on the claim that tracing in 2D is much easier than in 3D. To substantiate it, we conducted a small user study involving 5 PhD students used to performing such delineation tasks for research purposes. We asked them to annotate three volumes from the axon dataset using the Fiji Simple Neurite Tracer plugin [2], both in 2D and in 3D, and recorded how long it took them to complete these two tasks. We report the results in Fig. 4. For the two smaller volumes—292 × 292 × 40 and 231 × 231 × 71—it took people 3 to 4 minutes to create the 3D annotations and about 25% to 50% less in 2D. For the larger 335 × 335 × 67 volume, the 3D annotation time grew substantially but, it took about half as long to annotate in 2D.

While this study is too small to be statistically significant, it shows a clear trend: The larger the volume to be annotated, the more tedious the 3D annotation process, and the more attractive it becomes to annotate solely in 2D.

4.3 3D vs 2D Annotations

The 2D annotations are faster and easier but are *a priori* less informative than the 3D ones, and we could expect a performance drop when using the former. We now show that our framework *prevents* this drop from materializing.

	F1 score			Time saving [%]
	Axons	Retina	Angiography	
UNet/3D annot.	75.4	**81.5**	**77.6**	0
UNet/3 MIP annot./volume	**78.1**	78.2	75.9	50
UNet/2 MIP annot./volume	75.0	77.8	74.8	60
UNet/1 MIP annot./volume	72.3	39.0	57.7	70
Slice annot. [11]	70.8	75.8	74.1	50
Tubularity Score [4]	58.8	77.1	22.7	100
Centerline Detection [7]	68.5	62.6	50.3	0

Table 1. F1 score performance and corresponding time savings.

In Table 1, we compare the results obtained by training either on 2D or on 3D annotations in terms of the F1 score—the harmonic mean of precision and recall, which is a standard measure of binary segmentation performance—computed in 3D with respect to the 3D annotations. To ensure that the scores are comparable in both scenarios, we use the projections of the 3D annotations as our 2D annotations. In the rightmost column, we give an estimate, based on the above user study, of the time that could be saved by generating the 2D annotations instead of the 3D ones. In short, we obtain roughly the same performance—slightly better for the axons, slightly worse for the retina and brain scans—at half the annotation cost.

We can further reduce the amount of annotations used by training our approach using only 2 or even 1 single projection. The performance remains competitive when two projections are used, but decreases for a single one.

Whether using 3D or 2D annotations, these results rely on the modified U-Net architecture discussed in Section 3.3. For completeness, we also list in Table 1 the performance of an earlier Deep Net approach that relies on annotating a subset of slices [11]—and requires about the same amount of annotation as ours— and the performance attained by two older techniques [4, 7], which our approach also outperforms.

5 Conclusion

We have proposed a method for training DNNs to segment 3D images of linear structures using only annotations of 2D maximum intensity projections of the training data instead of full 3D annotations. We demonstrated that this results in decreased annotation requirements without loss of performance. To this end, we have exploited properties of visual hulls that are not specific to linear structures. In future work, we therefore intend to show that the scope of this technique is in fact much broader, for example by applying it to 3D membrane extraction.

References

1. Kutulakos, K., Seitz, S.: A Theory of Shape by Space Carving. IJCV **38**(3) (July 2000) 197–216
2. Frangi, A., Niessen, W., Vincken, K., Viergever, M.: Multiscale Vessel Enhancement Filtering. Lecture Notes in Computer Science **1496** (1998) 130–137
3. Law, M., Chung, A.: Three Dimensional Curvilinear Structure Detection Using Optimally Oriented Flux. In: ECCV. (2008)
4. Turetken, E., Becker, C., Glowacki, P., Benmansour, F., Fua, P.: Detecting Irregular Curvilinear Structures in Gray Scale and Color Imagery Using Multi-Directional Oriented Flux. In: ICCV. (December 2013)
5. Becker, C., Rigamonti, R., Lepetit, V., Fua, P.: Supervised Feature Learning for Curvilinear Structure Segmentation. In: MICCAI. (September 2013)
6. Breitenreicher, D., Sofka, M., Britzen, S., Zhou, S.: Hierarchical Discriminative Framework for Detecting Tubular Structures in 3D Images. In: MICCAI. (2013)
7. Sironi, A., Turetken, E., Lepetit, V., Fua, P.: Multiscale Centerline Detection. PAMI **38**(7) (2016) 1327–1341
8. Peng, H., Zhou, Z., E.Meijering, et al.: Automatic tracing of ultra-volumes of neuronal images. Nature Methods **14** (2017) 332–333
9. Peng, H., Tang, J., Xiao, H., et al.: Virtual Finger Boosts Three-Dimensional Imaging and Microsurgery as Well as Terabyte Volume Image Visualization and Analysis. Nature Communications **5** (2014) 4342
10. Vitanovski, D., Schaller, C., Hahn, D., Daum, V., Hornegger, J.: 3D Annotation and Manipulation of Medical Anatomical Structures. In: Proceedings of SPIE on Medical Imaging. Volume 7261. (2009)
11. Çiçek, Ö., Abdulkadir, A., Lienkamp, S., Brox, T., Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: MICCAI. (2016) 424–432
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI. (2015)
13. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv Preprint (2014)
14. Bullitt, E., Zeng, D., Gerig, G., et al.: Vessel Tortuosity and Brain Tumor Malignancy: A Blinded Study. Acad Radiol **12**(10) (October 2005) 1232–1240