SCONet: Convolutional Occupancy Networks for Multi-Organ Segmentation

Maylis Jouvencel, Razmig Kéchichian, Julie Digne, Sébastien Valette

SCONET: CONVOLUTIONAL OCCUPANCY NETWORKS FOR MULTI-ORGAN SEGMENTATION

Maylis Jouvencel[1] *Razmig Kéchichian*[1] *Julie Digne*[2] *Sébastien Valette*[1]

[1] INSA-Lyon, UCBL, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France
[2] Univ Lyon, INSA-Lyon, CNRS, UCBL, Centrale Lyon, Univ Lyon 2, LIRIS UMR 5205, F-69621, Lyon, France

ABSTRACT

Convolutional neural networks are the de facto standard for 3D multi-organ segmentation but still exhibit significant limitations, especially regarding their computational cost, with high running time and often prohibitive memory footprint for large 3D volumes. To overcome these limitations, we propose to replace the image voxel grid with a more compact point cloud representation. Recently, in the field of 3D object reconstruction, networks learning implicit functions from an input point cloud, such as Convolutional Occupancy Networks (ConvONet), have proven their good surface representation capabilities. We therefore propose SCONet (Segmentation Convolutional Occupancy Network), a lightweight ConvONet-based network adapted to the specific task of multi-organ segmentation. SCONet takes as input a point cloud extracted from the original volume with a standard contour detection algorithm, and enriches it with geometric and photometric features. Thanks to its ability to query per organ occupancy probabilities for any point in space, SCONet can be used to predict a multi-organ segmentation map at arbitrary resolution. We evaluate our method on an abdominal CT image dataset and compare its performances with those of discrete and implicit baselines. Our implementation is available at `https://github.com/maylis-j/SCONet`.

***Index Terms*—** Multi-organ segmentation, Point cloud, Implicit representation

1. INTRODUCTION

Abdominal organ segmentation is an important step for computer-aided diagnosis or treatment planning. However, manually providing segmentation maps is a very tedious task, which can now be efficiently automated by deep learning strategies. Recent methods achieved impressive results using Convolutional Neural Networks (CNNs), among which UNet has become the standard for the segmentation task [1, 2]. However CNNs are limited to grid-like input [3], and their computation cost can be excessive with heavy memory usage and often slow inference times as CT-scans can go up to 10^8 voxels [4]. To alleviate these problems, we propose to work on a point cloud extracted from the original image volume, a lighter representation allowing us to reduce the computational cost significantly. The use of point clouds in medical image segmentation has indeed gained attention recently [5, 4], but the solutions proposed often consist in simply adding a point cloud-based module to refine a CNN-produced segmentation which does not always guarantee lower computation cost Point clouds are also used as input for networks learning an implicit neural representation (INR). These networks usually learn a distance function, like DeepSDF [6], or an occupancy probability function, like Occupancy Networks (ONet) [7] and ConvONet [8] which can represent complex shapes at arbitrary resolutions through their indicator function defined in the continuous ambient space. Such networks have already been used for medical applications on tasks like image reconstruction [9] or segmentation [3]. In particular, methods using ONet with a CNN-based encoder showed promising results [10, 11, 12, 13] and confirm the interest of INRs for medical image segmentation from the perspective of low computation cost and good representation power. In this work, we propose to leverage a more recent ConvONet-based approach [8] with a greater local representation capacity thanks to a learned 32^3 feature grid which allows to learn more complex structures than the 512 dimensional latent vector from the original ONet architecture. Moreover, instead of processing the full volume directly, our network takes as input a point cloud which is extracted from the contours in the volume, and enriched with local photometric features.

The contributions of this paper are the following: (1) A multi-organ segmentation workflow based on a point cloud representation which can output segmentation maps at arbitrary resolution; (2) SCONet, a ConvONet-based network, designed to reconstruct several 3D objects from an input point cloud enriched with photometric and geometric features. We compare our method with voxel and INR networks on the task of abdominal multi-organ segmentation showing the advantages of our method in terms of memory usage and computational complexity while producing accurate results. We furthermore perform an ablation study which justifies our design choices.

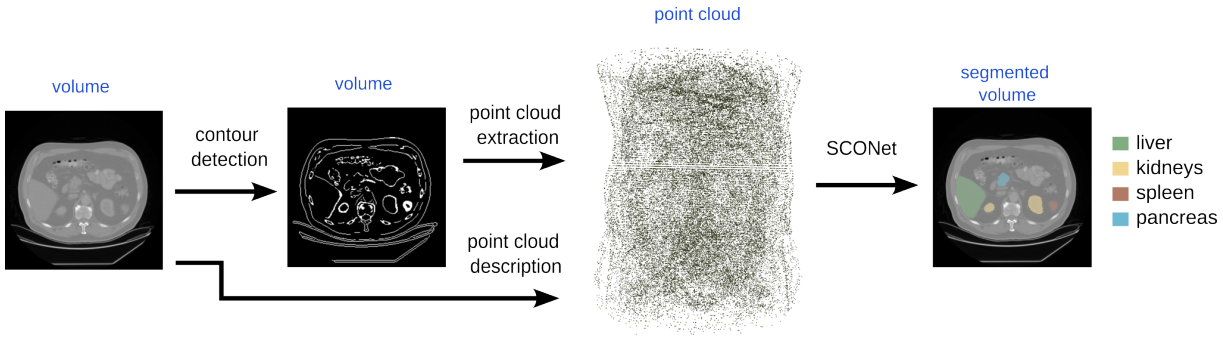

Fig. 1: Our full segmentation workflow for 3D images.

2. PROPOSED METHOD

2.1. Point cloud extraction

The first step of our workflow, illustrated in Figure 1, consists in contour point cloud extraction. We use the Canny edge detector [14] to extract a contour-based point cloud. For each point i of the point cloud, in addition to its coordinates (x_i, y_i, z_i), we compute the corresponding SURF descriptor which encodes local information in a 10^3-voxel neighbourhood around the point. This descriptor is computed with the 3D implementation [15] of the original algorithm [16] and yields a 48-valued vector $(s_i^1, ..., s_i^{48})$ for each point i. The resulting point cloud is a $(N \times 51)$ matrix which encodes geometric and local photometric features, where N is the number of points in the cloud. We translate and scale the (x_i, y_i, z_i) coordinates isotropically to fit in the unit cube.

2.2. SCONet

To process this point cloud, we introduce SCONet, a ConvONet-based network adapted from [8] to the specific task of multi-organ segmentation. The proposed network is displayed in Figure 2 and has the following encoder-decoder architecture.

Encoder: The encoder computes a feature grid from the contour-based input point cloud. Similarly to the original ConvONet, the coordinates of the input point cloud are fed to a PointNet encoder [17] which consists of 5 fully-connected ResNet blocks. This computes per point geometric features of dimension 32 which are then concatenated to the SURF descriptors of dimension 48. Features are then projected on a 32^3 grid by max pooling on each grid cell, resulting in a feature grid of size $(32^3 \times 80)$. This grid is then processed by a 3D U-Net with 3 levels which outputs a learned feature grid of dimension $32^3 \times 32$.

Decoder: The decoder uses the learned feature grid to predict the occupancy probabilities of a query point. A feature vector of size 32 is first computed through trilinear interpolation on the feature grid learned by the encoder. This feature vector is then fed to a ResNet-based Occupancy Network (ONet) [18] followed by a linear layer which outputs logits for all the classes. A final Softmax layer then outputs occupancy probabilities for the queried point for each organ.

2.3. Training and inference

Training: At training time we sample random query points in the 3D space and predict their occupancy values, optimizing a combined Dice [19] and cross-entropy loss.

Inference: During inference, we predict the segmentation map for the entire volume by querying points corresponding to the coordinates of all voxels in the original volume.

3. EXPERIMENTS

3.1. Dataset

We evaluate all methods on the AbdomenCT-1K dataset [20], which consists of 1000 abdominal CT scans from 12 medical centers. Image volumes have a resolution of $(512 \times 512 \times z)$, voxels with varying size in dimension z and varying sampling along the three axes. To achieve reasonable training times for voxel-based comparison baselines despite the large dataset and image sizes, we resample the original images along the x and y axes to obtain volumes of size $(256 \times 256 \times z)$. We keep the original sampling along the z axis to avoid possible degradation of results. Indeed the spacing along this axis often reaches 5 mm, which is much larger than the spacing along axes x and y. Reference segmentation maps are provided for the liver, kidneys, spleen and pancreas. We use a custom data split train/val/test of 806/91/100 images.

3.2. Training details

We implement SCONet with PyTorch and perform all experiments on a NVIDIA Tesla V100-SXM2-32GB GPU. We train the network during 300 epochs with the AdamW optimizer and a learning rate of 10^{-4}. Due to memory constraints, the mini-batch size is set to one input point cloud representing one subject. During training, the number of query points is set to $50k$. We also set the size of the input point cloud to $N = 50k$ for training and inference.

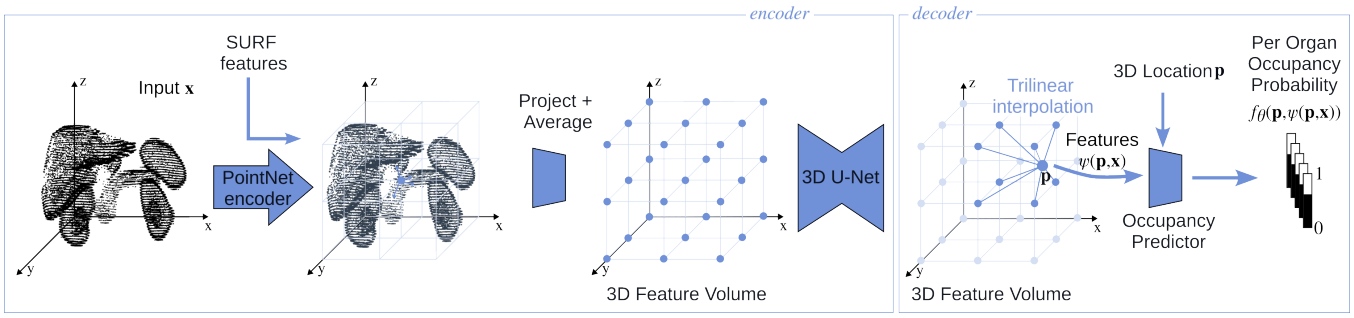

Fig. 2: SCONet workflow. For clarity purposes, the input point cloud displayed here shows only the contour points which belong to the target organs.

3.3. Comparison with other methods

To evaluate the advantages of working with a lighter point cloud representation and an implicit network, we compare our performances with both discrete and implicit methods. We choose 2D and 3D U-Net as CNN baselines since U-Net remains the backbone to many state-of-the-art methods. For both networks, we implement a 4-level architecture that we train with the Dice loss during 30 epochs. Mini-batch sizes used are 64 and 2 for 2D U-Net and 3D U-Net, respectively. Moreover we train the 3D U-Net on image patches of size $(256, 256, 30)$. We perform data augmentation in the training of both networks with with flips, rotations and elastic deformations. Our 2D and 3D U-Nets are based on the PyTorch implementation of [2] ([21]). SwinUNETR-V2 [22] serves as a transformer baseline, and is trained for 100 epochs with the weighted combination of cross-entropy loss and Dice loss. We also compare our network with ImPulSe [12], an INR baseline with a CNN encoder and ONet decoder, training it for 50 epochs with the weighted combination of cross-entropy loss and Dice loss from the original paper.

4. RESULTS

The first criterion to evaluate our method is the quality of predicted segmentation maps. Qualitative results presented in Figure 3 confirm that SCONet can produce meaningful segmentation maps, with an overall good segmentation of the four abdominal organs. Compared to segmentations predicted by the discrete baselines 2D and 3D U-Nets and SwinUNETR-V2, which could admittedly be improved through post-processing, both INR-based methods, SCONet and ImPulSe, directly produce a less noisy segmentation. However, SCONet has difficulties producing a high level of details with sometimes imprecise and over-smoothed borders, in particular for the kidneys.

Results of quantitative evaluation and comparison according to the Dice coefficient and the Hausdorff Distance (HD) are reported in Table 1. Bold face indicates significant performance differences with other methods according to the Wilcoxon test. We also study the influence of the feature grid size for SCONet: (A) corresponds to 32^3 and (B) to 64^3. Dice scores in Table 1 confirm the visual observation that SCONet achieves a lower segmentation accuracy compared to discrete methods, with SwinUNETR-V2 achieving best results. The pancreas seems particularly difficult to segment compared to the other organs. This can be explained not only by the high diversity of shapes for this organ but also by its local texture which results in poorly defined contours in the extracted point clouds, making the reconstruction more difficult in comparison to the three other organs. Although not as good as U-Nets and SwinUNETR-V2, ImPulSe yields better overall Dice scores compared to SCONet. However, the performance gap is reduced when we increase the size of the feature grid. From the viewpoint of segmentation surface precision, INR-based methods produce better HD, with SCONet having the best results. The lower HD from discrete strategies can be explained by their restriction to working on slices or patches, while both SCONet and ImPulSe process the full structure of the original volume.

We also report the computation cost of all compared methods in Table 1. In terms of GPU memory usage and number of parameters, our network is on par with 2D U-Net and 4 times more efficient than 3D U-Net. We are also more than 10 times more memory efficient compared to ImPulSe and SwinUNETR-V2. Note that SCONet can output segmentation maps with the same memory efficiency at any resolution thanks to its decoder based on query points. To evaluate computational complexity, the number of FLOPs is reported for a fixed volume size of $(256 \times 256 \times 100)$. We find that SCONet is the lightest network, with 4 times less GFLOPs than 2D U-Net for both configurations tested. We also note that the number of operations required by SCONet depends mostly on the number of query points since the encoder has a fixed number of operations, which is 15 GFLOPs for configuration (A) and 113 GFLOPs for (B).

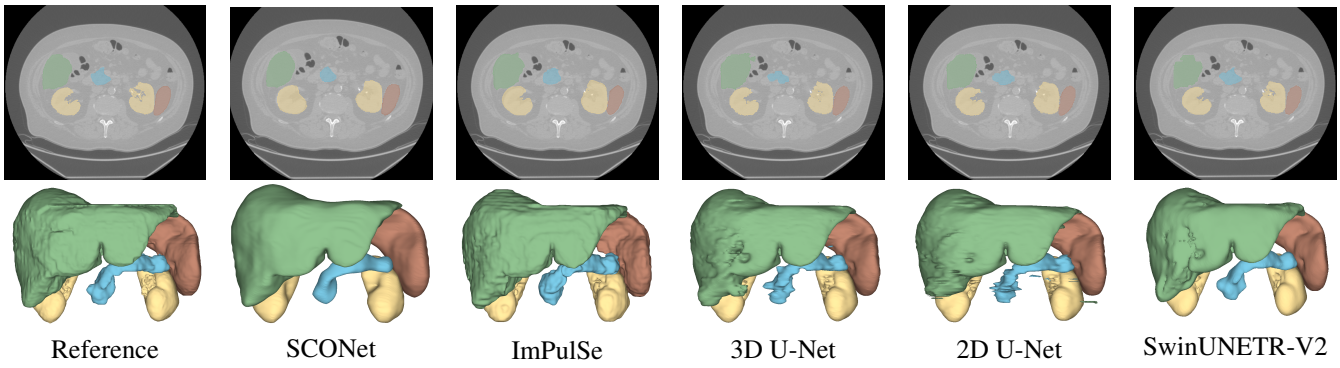

Fig. 3: Comparison of segmentation results obtained by SCONet (A), ImPulSe, 3D and 2D U-Net and SwinUNETR-V2 on an axial CT slice (top) and in 3D (bottom) for the liver (green), kidneys (yellow), pancreas (blue) and spleen (red).

Table 1. Performance comparison with the baselines. The version (A) of SCONet learns a feature grid with size 32^3, while the grid size for version (B) is 64^3

Model	**Dice**					**HD**	**Inf. GPU**	**#Param**	**GFLOPs**
	Liver	*Kidn.*	*Spleen*	*Panc.*	Avg.	Avg.	**(GB)**		
2D U-Net	0.961	0.943	0.950	0.794	0.912	91.60	**0.77**	1.36M	960
3D U-Net	0.956	0.938	0.948	0.808	0.913	46.29	4.48	4.08M	2,980
SwinUNETR-V2	0.950	0.936	0.954	0.830	**0.918**	139.16	7.24	18.35M	2,961
ImPulSe	0.957	0.929	0.930	0.763	0.895	20.32	16.25	33.28M	2,504
SCONet (A)	0.933	0.893	0.918	0.725	0.867	19.60	0.83	**1.26M**	**120**
SCONet (B)	0.934	0.904	0.924	0.769	0.882	**17.82**	1.27	**1.26M**	218

4.1. Ablation study

Table 2 presents an ablation study on the input features we use to enrich point clouds. We compare SURF features with intensity features, which are simply per-voxel intensity values in the original volume at the coordinates of each point. In addition, we compare these results with a version without any additional features, much like the original ConvONet which uses point coordinates only.

Table 2. SCONet segmentation results with different features.

Input features	**Dice** Avg	**HD** Avg	**Inf. GPU (GB)**	**#Param**	**GFLOPs**
None	0.788	26.09	0.78	1.07M	114
Intensity	0.820	23.82	0.78	1.07M	114
SURF	**0.867**	**19.60**	0.83	1.26M	120

Results reported in Table 2 confirm that enriching points with photometric features improves the segmentation quality significantly with only a small loss in computation efficiency. In particular, using SURF features improves both Dice and HD compared to using only voxel intensities. Indeed, the 48-valued SURF descriptor contains rich information about a point's neighborhood since it is computed on a 10^3 cube centered on the point. We note that feeding the SURF features directly to the PointNet encoder did not bring any improvements to the results.

5. CONCLUSIONS AND FUTURE WORK

We introduced SCONet, a lightweight ConvONet-based network which outputs a segmentation map at arbitrary resolution from a point cloud enriched with photometric features. Compared to discrete baselines as well as a similar INR-based network, SCONet shows a good trade-off between high segmentation quality and low computation cost, with a competitive memory usage and particularly low computational complexity, which is useful with limited GPU resources.

Our experiments show that SCONet segmentation results depend on the quality of the input point cloud. Missing points on the borders of organs can decrease the segmentation performance. This could be addressed by using a learning-based point extraction and by assessing the discrepancies between segmented organs boundaries and local point density. As a future work, we would also like to explore the utility of this method for multi-modality since preliminary results indicate good generalization of a network trained on CT data to MRI.

6. ACKNOWLEDGMENTS

This work was funded by the TOPACS ANR-19-CE45-0015 project of the French National Research Agency (ANR). It uses HPC resources from GENCI-IDRIS (Grant 2024-AD011013983R1).

7. REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI 2015*. Springer, 2015, pp. 234–241.

[2] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *MICCAI 2016*. Springer, 2016, pp. 424–432.

[3] N. Stolt-Ansó, J. McGinnis, J. Pan, K. Hammernik, and D. Rueckert, "Nisf: Neural implicit segmentation functions," in *MICCAI 2023*, 2023, pp. 734–744.

[4] N.-V. Ho, T. Nguyen, G.-H. Diep, N. Le, and B.-S. Hua, "Point-unet: A context-aware point-based neural network for volumetric segmentation," in *MICCAI 2021*. Springer, 2021, pp. 644–655.

[5] M. Ye, Q. Huang, D. Yang, P. Wu, J. Yi, L. Axel, and D. Metaxas, "Pc-u net: Learning to jointly reconstruct and segment the cardiac walls in 3d from ct data," in *Statistical Atlases and Computational Models of the Heart*. Springer, 2021, pp. 117–126.

[6] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *2019 IEEE CVPR*, 2019, pp. 165–174.

[7] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *2019 IEEE CVPR*, 2019, pp. 4460–4470.

[8] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *ECCV 2020*. Springer, 2020, pp. 523–540.

[9] A. Corona-Figueroa, J. Frawley, S. Bond-Taylor, S. Bethapudi, H. P. Shum, and C. G. Willcocks, "Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 3843–3848.

[10] C. Reich, T. Prangemeier, Ö. Cetin, and H. Koeppl, "OSS-Net: Memory Efficient High Resolution Semantic Segmentation of 3D Medical Data," in *British Machine Vision Conference*, 2021.

[11] M. O. Khan and Y. Fang, "Implicit neural representations for medical imaging segmentation," in *MICCAI 2022*. Springer, 2022, pp. 433–443.

[12] K. Kuang, L. Zhang, J. Li, H. Li, J. Chen, B. Du, and J. Yang, "What makes for automatic reconstruction of pulmonary segments," in *MICCAI 2022*. Springer, 2022, pp. 495–505.

[13] Y. Zhang, P. Gu, N. Sapkota, and D. Z. Chen, "Swipe: Efficient and robust medical image segmentation with implicit patch embeddings," in *MICCAI 2023*. Springer, 2023, pp. 315–326.

[14] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, , no. 6, pp. 679–698, 1986.

[15] R. Agier, S. Valette, R. Kéchichian, L. Fanton, and R. Prost, "Hubless keypoint-based 3d deformable groupwise registration," *Medical image analysis*, vol. 59, pp. 101564, 2020.

[16] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *ECCV 2006*. Springer, 2006, pp. 404–417.

[17] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE CVPR*, 2017, pp. 652–660.

[18] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *2020 IEEE CVPR*, 2020, pp. 3504–3515.

[19] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[20] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, et al., "Abdomenct-1k: Is abdominal organ segmentation a solved problem?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6695–6714, 2021.

[21] A. Wolny, L. Cerrone, A. Vijayan, R. Tofanelli, A. V. Barro, M. Louveaux, et al., "Accurate and versatile 3d segmentation of plant tissues at cellular resolution," *eLife*, vol. 9, pp. e57613, jul 2020.

[22] Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, and D. Xu, "Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation," in *MICCAI 2023*, 2023, pp. 416–426.