

# A LOCATION-SENSITIVE LOCAL PROTOTYPE NETWORK FOR FEW-SHOT MEDICAL IMAGE SEGMENTATION

*Qinji Yu<sup>1,\*</sup>, Kang Dang<sup>2,\*</sup>, Nima Tajbakhsh<sup>2</sup>, Demetri Terzopoulos<sup>3,2</sup>, Xiaowei Ding<sup>1,2</sup>*

<sup>1</sup> Shanghai Jiao Tong University, China

<sup>2</sup> VoxelCloud, Inc., Los Angeles, USA

<sup>3</sup> University of California, Los Angeles, USA

## ABSTRACT

Despite the tremendous success of deep neural networks in medical image segmentation, they typically require a large amount of costly, expert-level annotated data. Few-shot segmentation approaches address this issue by learning to transfer knowledge from limited quantities of labeled examples. Incorporating appropriate prior knowledge is critical in designing high-performance few-shot segmentation algorithms. Since strong spatial priors exist in many medical imaging modalities, we propose a prototype-based method—namely, the location-sensitive local prototype network—that leverages spatial priors to perform few-shot medical image segmentation. Our approach divides the difficult problem of segmenting the entire image with global prototypes into easily solvable subproblems of local region segmentation with local prototypes. For organ segmentation experiments on the VISCEAL CT image dataset, our method outperforms the state-of-the-art approaches by 10% in the mean Dice coefficient. Extensive ablation studies demonstrate the substantial benefits of incorporating spatial information and confirm the effectiveness of our approach.

**Index Terms**— medical image segmentation, few-shot segmentation, prototype networks, spatial layout priors

## 1. INTRODUCTION

The inception of deep neural networks has revolutionized the landscape of medical image segmentation [1, 2, 3]. Despite their tremendous success, however, depriving them of sufficient quantities of labeled images limits their effectiveness. Unfortunately, obtaining abundant amounts of expert-level-accurate, pixel-wise annotated data in the medical imaging domain incurs significant time and expense [4]. Therefore, it is imperative to explore new medical image segmentation techniques that can learn from scarcely annotated data.

Few-shot semantic segmentation is a relatively new trend in computer vision that addresses the data scarcity issue [5, 6, 7, 8, 9, 10]. A few-shot segmentation system learns how to transfer knowledge from a few labeled samples, typically referred to as support images, to guide image segmentation with novel classes, known as queries. The concept is motivated by human learning—the process of rapidly learning new

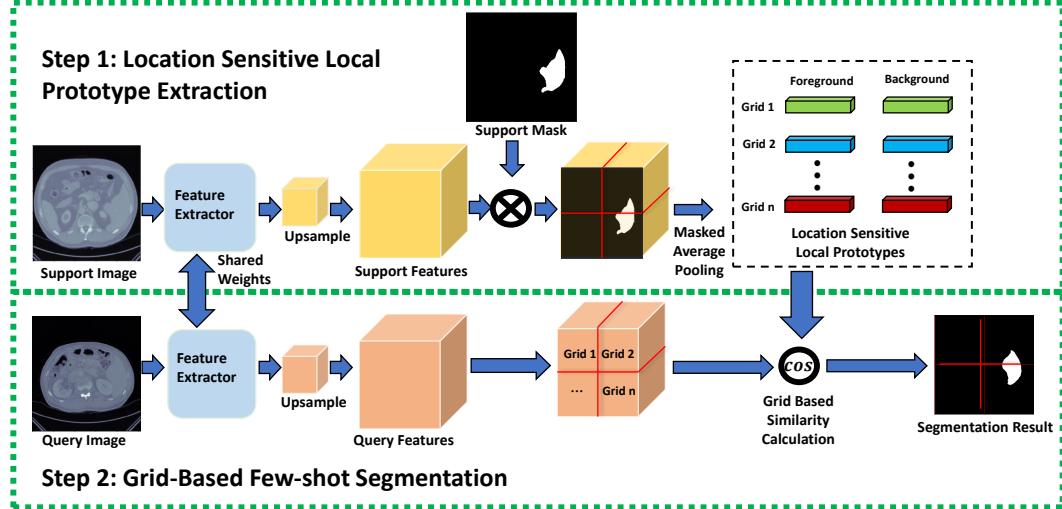
knowledge from a few observations by leveraging strong prior knowledge acquired from experience. Analogously, incorporating appropriate prior knowledge is critical in the development of high-performance few-shot semantic segmentation algorithms. It is well-known that strong spatial priors are inherent in many medical imaging modalities, including abdominal CT scans [11, 12]. For instance, the liver typically appears on the left side of the CT image slice, while the spleen appears on the right. Such spatial layout priors convey crucial information in the pixel-wise labeling of certain organ classes. This motivates us to study how to exploit spatial layout information within a few-shot image segmentation algorithm. To our knowledge, ours is the first attempt at incorporating spatial layout information in a few-shot medical image segmentation application.

In particular, our approach is a prototype-based method [7, 8] that we call the *location-sensitive local prototype network*. We divide the image into overlapping image grids and extract location-specific local prototypes for all the support image grids. Unlike the global prototypes used in the aforementioned publications, our local prototype is constrained to utilize the information within the corresponding grid; hence, it is location sensitive. Assuming that the support and query images have a similar spatial layout, each location in the query image can be matched to the corresponding local prototypes for pixel-wise segmentation. In this way, we divide the difficult problem of segmenting the entire image with global prototypes into easily solvable subproblems of local region segmentation with local prototypes, which leads to significantly improved performance. Hence, our contributions are as follows:

- Our novel location-sensitive local prototype network incorporates spatial layout information into the few-shot medical image segmentation application and demonstrates the substantial benefits of incorporating such information.
- We thoroughly analyze the grid size and other critical parameters in our experiments.
- We establish a new state-of-the-art performance for organ segmentation on the VISCEAL contrast-enhanced CT image dataset [13] by showing that our method outperforms the currently best approach [10] by 10% in the mean Dice coefficient, thereby confirming the effectiveness of our proposed approach.

---

\* Contributed equally.



**Fig. 1.** Processing pipeline of our location-sensitive local prototype network.

## 2. RELATED WORK

Few-shot semantic segmentation is a recent research topic. Promising schemes have appeared in recent years [5, 6, 7, 8, 9, 10]. Some efforts have focused on the few-shot medical image segmentation application [14, 15, 10, 16]. Tajbakhsh *et al.* [17] have surveyed this topic. Roy *et al.* [10] adopted squeeze-excitation modules to effectuate the interaction between support and query images in order to perform few-shot organ segmentation. Many approaches are based on non-parametric metric learning frameworks, such as prototype networks [6, 7, 9]. Wang *et al.* [7] proposed a prototype alignment regularization strategy between the support and query to learn a better class-specific global prototype representation. Ouyang *et al.* [16] designed an adaptive local prototype pooling module for prototypical networks to solve the foreground-background imbalance problem in medical image segmentation. Our approach is also prototype-based; however, unlike previous work, we focus on incorporating spatial layout information and demonstrate it to be highly beneficial to few-shot medical image segmentation.

## 3. METHODOLOGY

We first introduce the few-shot segmentation problem in Sec. 3.1, then explain the two main steps of our location-sensitive local prototype network (Fig. 1): location-sensitive local prototype extraction in Sec. 3.2, followed by grid-based few-shot segmentation in Sec. 3.3.

### 3.1. Problem Setting

The problem is to train a model on a large labeled training dataset  $\mathcal{D}_{\text{train}}$  that can perform segmentation on a testing dataset  $\mathcal{D}_{\text{test}}$  with only a few annotated examples, where

the class set  $\mathcal{C}_{\text{train}} \in \mathcal{D}_{\text{train}}$  has no overlap with the class set  $\mathcal{C}_{\text{test}} \in \mathcal{D}_{\text{test}}$ ; *i.e.*,  $\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} = \emptyset$ . For example, given the training set with training classes  $\mathcal{C}_{\text{train}} = \{\text{lungs}, \text{heart}\}$ , we expect the trained model to be proficient at segmenting a different set of testing classes  $\mathcal{C}_{\text{test}} = \{\text{kidney}, \text{bones}\}$  with reference only to a few labeled images in  $\mathcal{D}_{\text{test}}$ .

Like previous research [7, 16], we adopt an episodic paradigm for the few-shot segmentation task. Given an  $n$ -way,  $k$ -shot task, during model training we sample a set of episodes  $\mathcal{T}_{\text{train}}$  from  $\mathcal{D}_{\text{train}}$ . Each episode in  $\mathcal{T}_{\text{train}}$  is composed of a support set  $\mathcal{S}$  and a query set  $\mathcal{Q}$ , and the model is trained to transfer segmentation cues from  $\mathcal{S}$  to  $\mathcal{Q}$  in a class-agnostic manner. The support set  $\mathcal{S} = \{(I_i^s, M_{i,c}^s)\}$  contains  $1 \leq i \leq k$  pairs of support images  $I_i^s$  and masks  $M_{i,c}^s$  corresponding to a set of  $1 \leq c \leq n$  randomly picked classes  $\mathcal{C} \in \mathcal{C}_{\text{train}}$ . The query set  $\mathcal{Q} = \{(I_j^q, M_{j,c}^q)\}$  contains  $1 \leq j \leq n_q$  different query pairs likewise sampled from  $\mathcal{D}_{\text{train}}$  with class set  $\mathcal{C}$ . As each episode randomly samples different semantic classes, the trained model is expected to be class-agnostic and generalize well.

After training the segmentation model, we evaluate its performance on  $\mathcal{D}_{\text{test}}$ . During model testing, a set of episodes  $\mathcal{T}_{\text{test}}$  is obtained in similar fashion from  $\mathcal{D}_{\text{test}}$ , where the masks are associated with testing classes  $\mathcal{C}_{\text{test}}$ . For each testing episode, the segmentation model is evaluated on the query set  $\mathcal{Q}$  given the support set  $\mathcal{S}$ .

### 3.2. Location-Sensitive Local Prototype Extraction

Consider a coarse set  $\mathcal{G} = \{g_m\}$  of  $1 \leq m \leq n_g$  overlapping image grids uniformly distributed across the  $w \times h$  sized image. Grid  $g_m$  of size  $\alpha w \times \alpha h$ , where  $\alpha$  denotes the grid scale, is centered at  $(x_m, y_m)$ . To extract location-sensitive local prototypes, support images are passed to the feature encoder to generate the corresponding features maps. The location-sensitive local prototype of  $g_m$  is calculated from support

images and masks by masked average pooling, as follows:

$$p_{c,g_m}^s = \frac{1}{k} \sum_{i=1}^k \frac{\sum_{(x,y) \in g_m} F_i^s(x,y) M_{i,c}^s(x,y)}{\sum_{(x,y) \in g_m} M_{i,c}^s(x,y)}, \quad (1)$$

where  $(x, y)$  are spatial coordinates and  $F_i^s$  are the extracted feature maps. If  $\sum_{(x,y) \in g_m} M_{i,c}^s(x,y) = 0$ , we simply set  $p_{c,g_m}^s = 0$ . We also treat the background as a semantic class whose prototype  $p_{0,g_m}^s$  is computed by performing masked average pooling on all locations within  $g_m$  excluding any foreground classes. The following are key observations regarding the extracted local prototype  $p_{c,g_m}^s$ :

1. Prototype  $p_{c,g_m}^s$  is associated with the local image grid  $g_m$  and, hence, is location sensitive. Grid scale  $\alpha$  is a critical parameter influencing the spatial information encoded by the local prototype. When  $\alpha$  is so small that the grid contains only a single pixel, the extracted local prototype becomes overly location sensitive, whereas when  $\alpha = 1$ , the grid expands to the full size of the image and position information is not encoded at all.
2. Unlike global prototypes, our local prototype  $p_{c,g_m}^s$  is constrained to utilize information only within the corresponding grid  $g_m$ . Hence, the loss of local information due to the averaging out of features outside the local grid can be prevented. This is essential for constructing high-quality background class prototypes  $p_{0,g_m}^s$ , because the background class tends to be spatially inhomogeneous, involving many different anatomical structures whose information would be unreasonably averaged out in the absence of the local grid constraint.

We collect the extracted local prototypes for all the grid locations and classes into set  $\mathcal{P} = \{p_{c,g_m}^s\}$ , where  $1 \leq m \leq n_g$  and  $0 \leq c \leq n$ .

### 3.3. Grid-Based Few-Shot Segmentation

Assuming that the support and query images have a similar spatial layout, local prototypes  $\mathcal{P}$  extracted from the support set can be matched to the corresponding locations in the query images in order to perform grid-based few-shot segmentation. Given the support prototype set  $\mathcal{P}$  of  $n$  foreground classes and the background class, the probability score of the query image corresponding to class  $c$  is obtained as

$$P_{j,c}^q(x,y) = \sigma(\text{sim}(F_j^q(x,y), \mathcal{P})), \quad (2)$$

where  $\text{sim}(F_j^q(x,y), \mathcal{P})$  is the similarity score between the query feature vector  $F_j^q(x,y)$  and the support prototype set  $\mathcal{P}$ , and  $\sigma$  is the softmax operation.

For non-overlapping image grids, we may employ the cosine similarity between the query feature vector  $F_j^q(x,y)$  and the corresponding matched grid prototype  $p_{c,g_m}^s$ , where the query feature vector location  $(x,y)$  belongs to the grid  $g_m$ . With overlapping image grids, however, the feature at location  $(x,y)$  may be mapped to multiple grids. We collect

Method	liver	spleen	kidney	psoas	Mean
PANet [7]	59.4	24.1	23.7	15.6	30.7
SENet [10]	70.0	60.7	46.4	49.9	56.7
Our Result	77.9	71.5	67.5	49.9	66.7
w/ Added Classes	<b>79.3</b>	<b>73.3</b>	<b>76.5</b>	<b>52.4</b>	<b>70.3</b>
ALPNet [16]	78.3	70.9	72.1	-	-

**Table 1.** Comparisons against state-of-the-art methods.

all the grids that influence location  $(x,y)$  into a set  $\Omega = \{g_m : (x,y) \in (x_m, y_m, \alpha)\}$  and calculate the similarity score as

$$\text{sim}(F_j^q(x,y), \mathcal{P}) = \max_{g_m \in \Omega} (\cos(F_j^q(x,y), p_{c,g_m}^s)), \quad (3)$$

where  $\cos$  denotes the cosine similarity function.

The network can be trained end-to-end using the standard cross-entropy loss:  $L_{ce}(P^q, M^q)$ , where  $P^q$  and  $M^q$  refer, respectively, to the predicted probability maps and ground-truth query masks of the particular training batch. For network evaluation, the predicted segmentation label of query image  $I_j^q$  at spatial coordinate  $(x,y)$  is then given by:  $\hat{M}_j^q(x,y) = \arg \max_c P_{j,c}^q(x,y)$ .

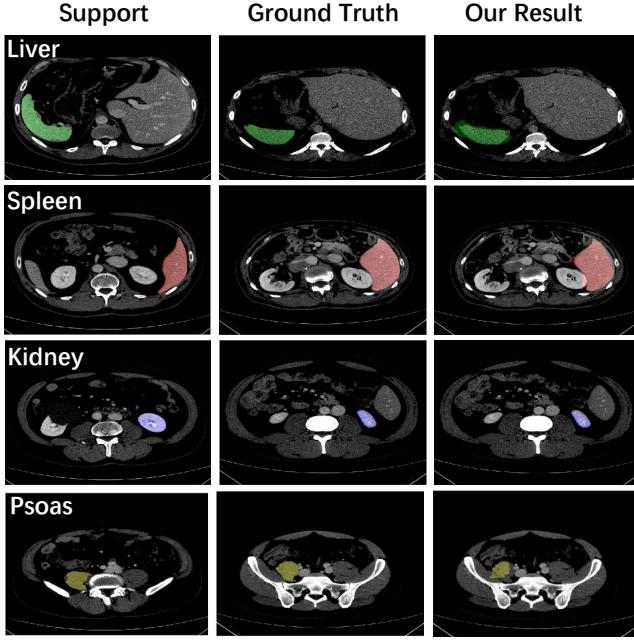
## 4. EXPERIMENTS

For all our experiments, we utilized contrast-enhanced CT scans from the VISCERAL dataset [13]. We used 65 silver corpus scans for training and 20 gold corpus scans for testing. We performed 1-way, 1-shot experiments for six organ classes: liver, spleen, left kidney, right kidney, left psoas, right psoas. For a fair comparison, most experiments reported in Sec. 4.1 were performed using the same dataset configuration used by Roy *et al.* [10]. In addition to the six organ classes, the VISCERAL dataset contains for each scan 14 additional organ annotations not used by these authors, which we use in other experiments. To evaluate 2D segmentation performance on 3D volumetric images, we follow the protocol described in [10] and use the Dice score as the metric.

Like [8], we use as a feature encoder a customized version of ResNet-50 [18] pre-trained on ImageNet [19]. CT slices are resized to  $512 \times 512$  before they are input to our network. We set the grid scale  $\alpha = 1/8$  and spacing between grid centers to half the grid size for most experiments. We apply RandomGamma, RandomContrast, and RandomBrightness to augment the training images. The models are trained by SGD [20] with a batch size of 4 for 10,000 iterations. The initial learning rate is  $10^{-3}$  and is decayed by 0.1 per 2,500 iterations.

### 4.1. Comparisons Against the State of the Art

Referring to Table 1, with an experimental protocol identical to that of the SENet few-shot organ segmentation model [10],

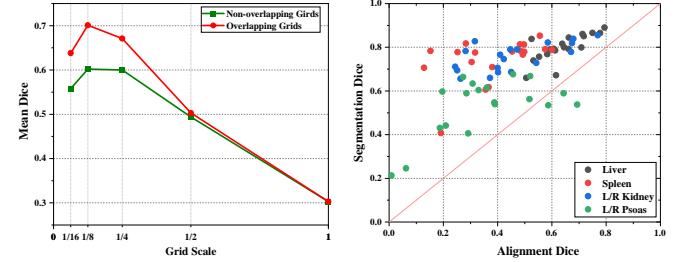


**Fig. 2.** Visual results. (Left) Support images and the associated ground-truth masks. (Middle) Query images and the associated ground-truth masks. (Right) Query images overlayed with our segmentation results.

our model achieved state-of-the-art performance in all the organ classes and a mean Dice coefficient of 66.7%, compared to SENet’s best-reported performance of 56.7%. PANet [7] is a popular prototype network previously tested only on natural image datasets; we experimented with its code on the VISCERAL dataset. Unlike our approach, both SENet and PANet do not explicitly incorporate a spatial layout prior. As is demonstrated by our ablation study, reported below, appropriately encoding spatial information is critical for good performance, and explains why our method performs significantly better than previous approaches.

By taking advantage of additional organ classes during model training, we obtained a better mean Dice score of 70.3%. The goal of few-shot segmentation is to construct a *class-agnostic* model by transferring semantic information from the support set to query images for unseen classes. Adding more class types increases training set diversity, prevents the trained model from overfitting to particular training classes, and produces better generalization-capable models. ALPNet [16] attains this goal with a self-supervised training technique for few-shot segmentation. We note that ALPNet achieves a similar Dice score to our method for some organs, although the experiments reported in [16] were conducted on different datasets, thus precluding a fair comparison.

Fig. 2 presents some segmentation visualizations. Evidently, our method successfully transfers semantic information from support to query images to produce high-quality segmentation results despite different organ sizes and shapes. Fig. 4 (on the last page of this paper) shows additional segmentation



**Fig. 3.** Ablation study. (Left) Benefits of selecting an appropriate grid scale and using overlapping image grids. (Right) Performance of our approach for different degrees of spatial misalignment.

results of our method compared with other approaches.

#### 4.2. Ablation Study

The left graph of Fig. 3 shows the benefit of selecting an appropriate grid scale. When the grid scale  $\alpha$  is set to 1/16 or 1/4 instead of the optimal value of 1/8, the location prior may be overly or insufficiently utilized and the mean Dice score drops from 70.3% to 63.8% or 67.1%, respectively. When  $\alpha = 1$ , the grid covers the entire image and the mean Dice score falls to 30.3%, because the location prior is not utilized at all. The resultant performance drop reveals the significance of location information. The graph also shows that when using non-overlapping rather than overlapping image grids, the mean Dice score decreases from 70.3% to 60.2%, thus confirming the benefit of overlapping image grids.

Our approach relies on the assumption that the support and query images have a similar layout and are somewhat spatially aligned. We evaluated its performance when this assumption fails. The right graph of Fig. 3 shows the performance of our approach for different degrees of spatial misalignment. From the testing set, for each support-query pair and each organ class, we calculated the Dice coefficient between the ground-truth support mask and the ground-truth query mask, indicated as “alignment Dice” on the horizontal axis of the graph. We also calculated the Dice score between the segmentation result and the ground-truth query mask, indicated as “segmentation Dice” on the vertical axis. The graph shows that our method is quite robust against spatial misalignments—for organs such as liver, spleen, and l/r kidney, many data points on the graph are associated with an alignment Dice of less than 40%. This implies that for the corresponding support-query pairs, the ground-truth support and query masks have little spatial overlap and, hence, are poorly aligned. Nevertheless, the segmentation Dice scores for these support-query pairs are often higher than 60%, signifying good performance despite the spatial misalignments between the ground-truth support and query masks. Our method performs less well for l/r psoas, as these are small anatomical structures whose visual appearance is similar to the surrounding structures, making it harder to learn discriminative visual features. Overall this

experiment demonstrated that our method has good misalignment tolerance for certain organs.

## 5. CONCLUSIONS

We proposed a prototype-based method that leverages spatial layout priors in performing few-shot medical image segmentation. Exploiting local prototypes, our novel location-sensitive local prototype network divides the hard problem of segmenting the entire image into easily solvable local region segmentation subproblems, yielding significantly improved performance. In organ segmentation experiments on the VISCERAL dataset, our model outperformed the current state of the art by 10% in the mean Dice coefficient. Our extensive ablation studies demonstrated the substantial benefits of incorporating spatial information.

## 6. COMPLIANCE WITH ETHICAL STANDARDS

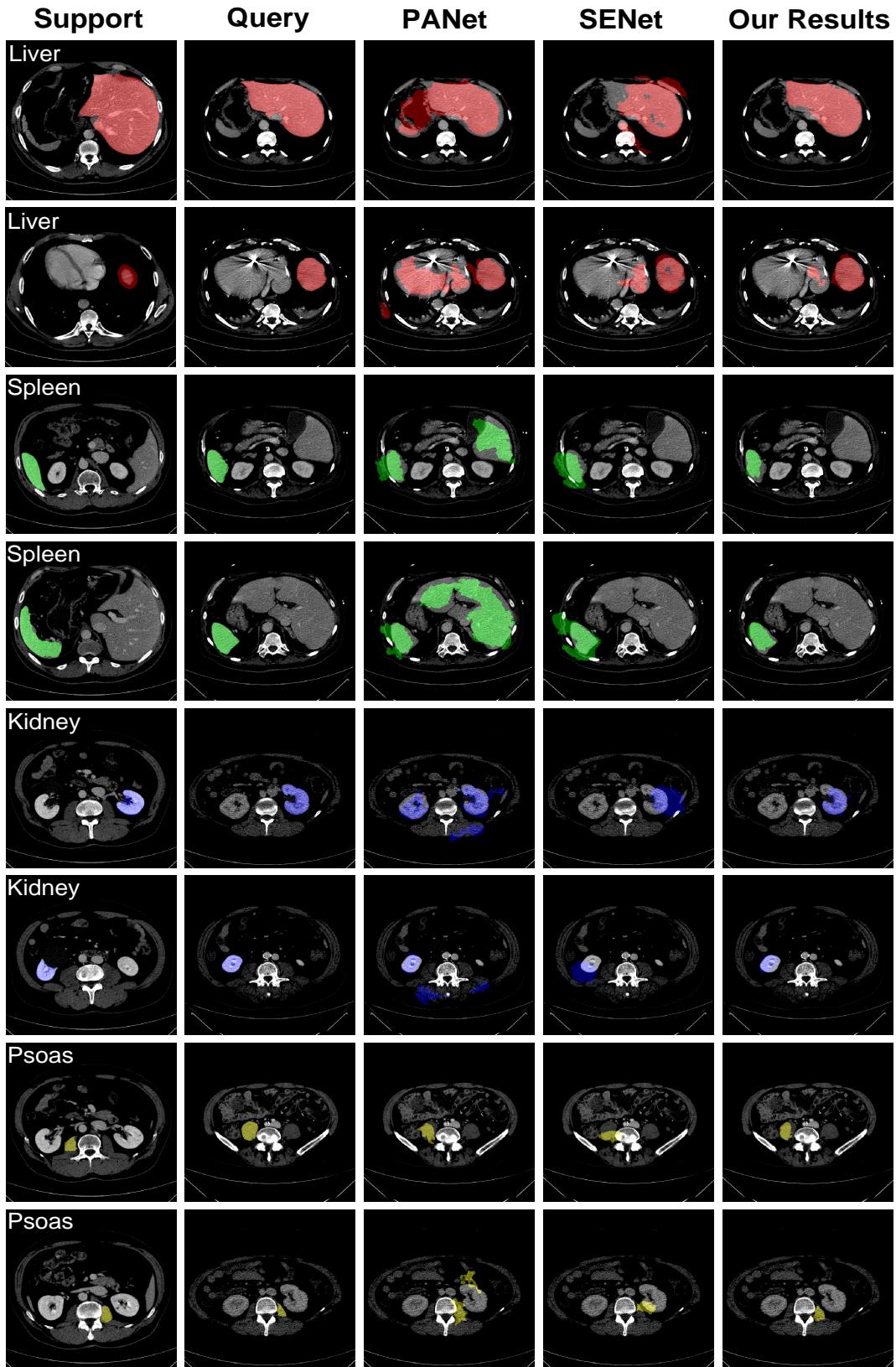
This research study was conducted retrospectively using human subject data of annotated anatomical structures made available in open access by the VISCERAL Consortium. Ethical approval was not required as confirmed by the license attached with the open access data.

## 7. ACKNOWLEDGMENTS

This research was supported by VoxelCloud, Inc. XD and DT are founders of the company.

## 8. REFERENCES

- [1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
- [2] F. Isensee, J. Petersen, A. Klein, et al., “nnU-Net: Self-adapting framework for U-Net-based medical image segmentation,” *arXiv preprint arXiv:1809.10486*, 2018.
- [3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.
- [4] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding, “Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data,” in *IEEE International Symposium on Biomedical Imaging*, 2019, pp. 1251–1255.
- [5] K. Rakelly, E. Shelhamer, T. Darrell, A.A. Efros, and S. Levine, “Few-shot segmentation propagation with guided networks,” *arXiv preprint arXiv:1806.07373*, 2018.
- [6] N. Dong and E.P. Xing, “Few-shot semantic segmentation with prototype learning,” in *British Machine Vision Conference*. 2018, p. 79, BMVA Press.
- [7] K. Wang, J.H. Liew, Y. Zou, D. Zhou, and J. Feng, “PANet: Few-shot image semantic segmentation with prototype alignment,” in *IEEE International Conference on Computer Vision*, 2019, pp. 9197–9206.
- [8] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, “CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.
- [9] J. Liu and Y. Qin, “Prototype refinement network for few-shot segmentation,” *arXiv preprint arXiv:2002.03579*, 2020.
- [10] A.G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, “‘Squeeze & excite’ guided few-shot segmentation of volumetric images,” *Medical Image Analysis*, vol. 59, pp. 101587, 2020.
- [11] K. Dang and J. Yuan, “Learning location constrained pixel classifiers for image parsing,” *Journal of Visual Communication and Image Representation*, vol. 49, pp. 1–13, 2017.
- [12] F. Haghghi, M.R. Hosseinzadeh Taher, Z. Zhou, M.B. Gotway, and J. Liang, “Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration,” in *Medical Image Computing and Computer Assisted Intervention*. 2020, pp. 137–147, Springer.
- [13] O. Jimenez-del Toro, H. Müller, M. Krenn, et al., “Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2459–2475, 2016.
- [14] A.K. Mondal, J. Dolz, and C. Desrosiers, “Few-shot 3D multi-modal medical image segmentation using generative adversarial learning,” *CoRR*, vol. abs/1810.12241, 2018.
- [15] A. Zhao, G. Balakrishnan, F. Durand, J.V. Guttag, and A.V. Dalca, “Data augmentation using learned transformations for one-shot medical image segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8543–8553.
- [16] C. Ouyang, C. Biffl, C. Chen, T. Kart, H. Qiu, and D. Rueckert, “Self-supervision with superpixels: Training few-shot medical image segmentation without annotation,” *arXiv preprint arXiv:2007.09886*, 2020.
- [17] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J.N. Chiang, Z. Wu, and X. Ding, “Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation,” *Medical Image Analysis*, p. 101693, 2020.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conf. on Computer Vision and Pattern Recog.*, 2016, pp. 770–778.
- [19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [20] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT 2010*, pp. 177–186. Springer, 2010.



**Fig. 4.** Additional visualizations. (Col.1) Support images and the associated ground-truth masks. (Col.2) Query images and the associated ground-truth masks. (Col.3) PANet results. (Col.4) SENet results. (Col.5) Our results.