Memory-based Few-Shot Medical Image Segmentation with Attention Mechanism

Yingxin Lai[1](31520221154209), Guoqing Yang[1](31520221154231), Rong Wang[1](36920221153120), Haoke Xiao[1](36920221153127)

[1]Class AI

Abstract

Recent work has shown that label-efficient few-shot learning through self-supervision can achieve promising medical image segmentation results. However, few-shot segmentation models typically rely on prototype representations of the semantic classes, resulting in a loss of local information that can degrade performance. This is particularly problematic for the typically large and highly heterogeneous background class in medical image segmentation problems. Previous works have attempted to address this issue by learning additional prototypes for each class, but since the prototypes are based on a limited number of slices, we argue that this ad-hoc solution is insufficient to capture the background properties. Motivated by this, we propose a novel memory architecture with an attention mechanism for few-shot medical image segmentation in which we refrain from modeling the background explicitly, and the triplet loss function is introduced to guide the model to learn more useful features from the image.

Introduction

Medical image segmentation is essential for auxiliary medical in medical image analysis (Minaee et al. 2020). It is used for lots of clinical practices and medical anatomy image studies, including tumor detection, disease diagnosis, and so on. Recently, deep learning has achieved high accuracy in segmentation tasks. However, such success of classical neural networks always requires a large amount of high-accuracy annotations, which are hard to be gotten in the real world due to the reliable clinical expertise, time-consuming and expensive cost associated with annotation.

In order to address these two challenges, few-shot learning has been proposed (Hu, Shen, and Sun 2018a) as one of the potential solutions. Most few-short learning methods adopt an episode-based meta-learning strategy. The dataset has been separated into two subsets: support set and query set. At train time, the same class image will be selected in those two subsets respectively in one episode for training. Models are expected to learn some instructional features in support images to guide the segmentation in query image (Wang et al. 2019). At test time, the model is expected to predict mask for query image with the guidance of the support set in a new class that does not appear in the training set.

It is a challenging task to make exact semantic segmentation in medical images with such extremely lack of annotated images (Siam et al. 2020).

Few-shot segmentation (FSS) In a recent work, based on prototypical FSS (Liu et al. 2020), SSLALP (Ouyang et al. 2020) propose a label-efficient approach to medical image segmentation. They proposed a novel algorithm that instead of sampling labeled support and query images and used superpixel serves to generate fake labels to train the model without annotations. But using fake masks to train the model does not consider the real characters of exact organs.

Furthermore, in medical image, considering the severe imbalance between foreground and background, it is not reasonable to do global average pooling and generate a few prototypes for background and foreground respectively (Lieb, Lookingbill, and Thrun 2005), which will miss a lot of local context features. To solve this problem, SSLALP proposed to use ALPNet to generate a series of prototypes adaptively to maintain more useful information (Snell, Swersky, and Zemel 2017). At test time, they only need a handful of labeled images to predict masks for query images on new classes. However, the effect is not significant, a general problem with SSLALP is the loss of local information caused by average pooling of features during prototype extraction (Vinyals et al. 2016). This is particularly problematic for spatially heterogeneous classes like the background class in medical image segmentation problems, which can contain any semantic class other than the foreground class (Mondal, Dolz, and Desrosiers 2018).

In order to solve the above problems[1], we introduce a new structure to keep the most discriminative prototypes using memory bank for each class respectively. The structure is shown in Fig.1. After recording the discriminative prototypes, we can guide the model learning by using prototypes stored in memory bank. At train time, according to mask provided we can get the anchor. The prototypes in memory bank with the similar label are positive samples and with the different label are negative samples. Meanwhile, in medical image segmentation, the local context feature is an important basis to determine the boundary between foreground and background. To enhance and emphasize these features,

---

[1]Code at https://github.com/laiyingxin2/Memory-based-Few-Shot-Medical-Image-Segmentation-with-Attention-Mechanism

in this paper we use attention mechanisms and propose the context-relational encoder to enhance the context features and forces the model to focus on the shape and context of the region of interest rather than the pixel itself. The main contributions of our work are as follows:

- Introduce the SE attention to explicitly model the interdependence between feature channels. By learning to automatically obtain the importance of each feature channel.
- Propose a novel memory architecture for medical images in order to store the global information of the whole dataset.
- Introduce Triplet loss function to guide model learning.
- Use abdominal MRI to prove the efficiency of our method. more useful features from images

Related Work

Medical Image Segmentation Based on U-Net

Medical image segmentation is the solution of clinical problems by analyzing images obtained by medical imaging systems. The purpose is to extract effective information and improve the level of clinical diagnosis (Liu et al. 2021). In recent years, image segmentation based on deep learning methods has been widely used. U-net (Ronneberger, Fischer, and Brox 2015) is one of the most important semantic segmentation frameworks for convolutional neural networks (CNN), which is proposed to apply end-to-end training to medical image analysis. Recently, the latest improved model of U-Net, nnUNet (Isensee et al. 2019), has implemented state-of-the-art in various datasets, which combines different network architectures to automatically configure optimal settings for different tasks. However, all these methods need to be trained on a large number of human annotations to achieve good performance, but the performance drops drastically when segmenting new classes.

SE Attention

SE Attention mechanism is used in SE-Net. It was proposed by Hu et al (Hu, Shen, and Sun 2018b). The attention mechanism guides computing resources to the most informative part of the input signal. It is generally used in combination with threshold functions (such as softmax and sigmoid) or sequence methods. SE block is a lightweight threshold mechanism specially used to model the correlation of each channel. One of its advantages is that it can effectively improve the attention mechanism and improve the accuracy of the network. In recent years, SE-Net has been widely used in many research fields. A researcher (Deng et al. 2022) propose an improved convolutional neural network framework that integrates innovative SE-Attention mechanism to learn discriminative features. (Gutiérrez-Viedma et al. 2021) aims to describe the precise relationship between management timing, duration, and prognosis of SE. The proposed method achieved an accuracy of 85.8% on the M2CAI16 workflow challenge dataset (Li et al. 2021).In this paper, we use his channel attention mechanism to achieve a more accurate distinction between the foreground and background feature maps.

Prototype Memory Mechanism

A memory bank can capture diverse features over a long time period and has been shown to be effective in other domains. (Hansen et al. 2022) utilize memory architecture for maintaining more useful features on the entire dataset and achieve good performance in other fields, such as person re-id. (Wu et al. 2018) use memory banks in comparative learning to obtain more diverse negative samples for comparison. (Caron et al. 2020) also use memory queues during training to accumulate representative negative samples. (He et al. 2022) show that using momentum update encoders can improve the stability of features accumulated in the memory bank, and we also use this strategy when learning prototypes. SimGAN (Shrivastava et al. 2016) introduce the image pool technology, which uses buffers to store previously generated samples so that the discriminator can not only focus on the current training batch but also self-improve based on memory.

Few-shot Semantic Segmentation

In recent years, few-shot segmentation has received more and more attention. (Shaban et al. 2017) first proposed a model for few-shot segmentation. Compared with classical semantic segmentation where models are trained with massive annotated data to perform one specific task well, few-shot semantic segmentation (Tang et al. 2021) requires models to be trained with a handful of annotated data in a train set and have a good performance in test set whose labels are new and unseen in training. (Roy et al. 2019) directly concatenates the support and query features as input to predict the masks for query images. In the prototype-based methods (Ouyang et al. 2020), the support images are used to generate prototypes guiding the segmentation of query images via a prototype-query comparison module. However, all these prototype-based methods are limited by the episodic training approach (Bienenstock, Geman, and Potter 1996), which ignores the overall representation across the entire dataset (He et al. 2022).

Method

In part 1, we will introduce the formal definition of few-shot medical image segmentation. In part 2, We introduce network architecture. And describes the main modules, the important two parts are SE attention and Memory Bank. In part 3, we employ Triplet loss to use the prior information stored in memory adequately to let model learn more differences between different organs.

Problem definition

Given a labeled dataset with classes $\mathbf{C}_{train}$ (e.g. $\mathbf{C}_{train} = \{liver, spleen\}$), FSS models are supposed to learn how to segment novel classes in $\mathbf{C}_{test}$ (e.g. $\mathbf{C}_{test} = \{left - kidney, right - kidney\}$) with a handful of labeled samples. $\mathbf{D}_{train} = \{(\mathbf{x}, \mathbf{y}(c^{\hat{j}}))\}$, where $\mathbf{x} \in \mathcal{X}$ is composed of images in train set and $\mathbf{y}(c^{\hat{j}}) \in \mathcal{Y}$ denotes the binary masks corresponding $\mathbf{x}$, where $\hat{j} = 1, 2, 3, \ldots, N$ is the class index. $\mathbf{D}_{test}$ is defined as the same as $\mathbf{D}_{train}$ except
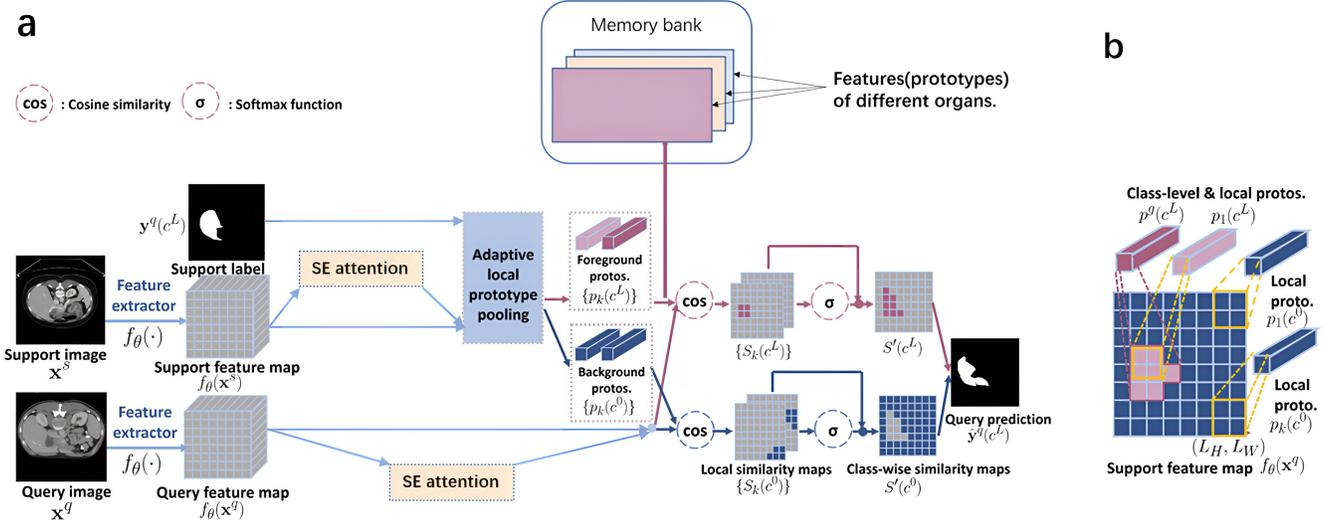
Figure 1: (a). Workflow of the proposed network: The feature extractor $f_\theta(\cdot)$ takes the support image and query image as input to generate feature maps $f_\theta(x^s)$ for support and $f_\theta(x^q)$ for query. Next, by introducing SE Attention learning to automatically capture the importance of each feature channel, enhance useful features and suppress features that are not useful for the current task. Then, Adaptive local prototype pooling module and takes support feature map and support label as input to obtain an ensemble of representation prototypes $p_k(c^j)'s$. These prototypes will be stored in Memory bank by their classes respectively, then will be used as references for comparing with query feature map $f_\theta(x^q)$.Similarity maps generated by these comparisons are fused together to form the final segmentation. This figure illustrates a 1-way segmentation setting, where c^L is the foreground class and c^0 is the background. (b). Illustration of the adaptive local prototype pooling module: Local prototypes are calculated by spatially averaging support feature maps within pooling windows (orange boxes); class-level prototypes are averaged under the entire support label (purple region).

the subscript. In each inference pass, a support set $\mathcal{S}$ and a query set $\mathcal{Q}$ are given. The support $\mathcal{S} = \{(\mathbf{x}_l^s, \mathbf{y}_l^s(c^{\hat{j}}))\}$ is composed of images $\mathbf{x}_l^s$ and its binary masks $\mathbf{y}_l^s(c^{\hat{j}})$. The query set $\mathcal{Q} = \{\mathbf{x}^q\}$ contains images $\mathbf{x}^q$ that are going to be segmented. Here, the superscripts denote an image or mask from support(s) or query(q). And $l = 1, 2, 3, \ldots, K$ is the index for each image-mask pair of class $c^{\hat{j}}$. One support-query pair $(\mathcal{S}, \mathcal{Q})$ forms an episode. In each episode, the Support-query pair will be sampled randomly.

Network Architecture

Overview Our network is composed of:(a) a generic feature extractor network $f_\theta(\cdot) : \mathcal{X} \to \xi$ parameterized by θ and the adaptive local prototype pooling module(ALP) $g(\cdot, \cdot): \xi \times \mathcal{Y} \to \xi$ for extracting representation prototypes from support features and labels. (b) SE Attention (Hu, Shen, and Sun 2018b) is added to support and query's feature map to explicitly model the interdependence feature between channels, and the importance of each feature channel is obtained by learning. (c)a similarity based classifier $sim(\cdot, \cdot) : \xi \times \xi \to \mathcal{Y}$ for segmentation by comparing query features and prototypes. (d) the proposed memory bank $M(\cdot)$ to maintain the prior information of each class in training set to guide model training.

Feature Extractor and Adaptive local pooling In inference, the feature extractor $f_\theta(\cdot)$ will extract representations

from support image and query image as $f_\theta(x^s)$ and $f_\theta(x^q)$ respectively. Our model will use gray histogram to extract features. By analyzing the gray level of the image, we can get the gray histogram of the image, from which we can extract many effective gray features.

Some researchers (Ouyang et al. 2020) propose to preserve local information in prototypes by introducing adaptive local prototype pooling module (ALP). In ALP, each local prototype is only computed within a local pooling window overlaid on the support and only represents one part of the object-of-interest. Specifically, The average pooling with a pooling window size (L_H, L_W) on each $f_\theta(\mathbf{x}_l^s) \in R^{D \times H \times W}$ where (H, W) is the spatial size and D is the channel depth. The obtained local prototype $p_{l,mn}(c)$ with undecided class c at spatial location (m, n) of the average-pooled feature map is given by Eq. (1).

$$p_{l,mn}(c) = avgpool\left(f_\theta\left(\mathbf{x}_l^s\right)\right)(m, n)$$

$$= \frac{1}{L_H L_W} \sum_h \sum_w f_\theta\left(\mathbf{x}_l^s\right)(h, w) \quad (1)$$

where $mL_H \le h < (m+1)L_H, nL_W \le w < (n+1)L_W$.

SE Attention After the feature extraction layer, the Squeeze-and-Excite attention (Hu, Shen, and Sun 2018b) module $SE(\cdot)$ is introduced, which consists of a global pooling layer, a fully connected layer, a ReLU activation layer, and a sigmoid activation layer. The SE block only increases a small amount of model complexity and computational cost,
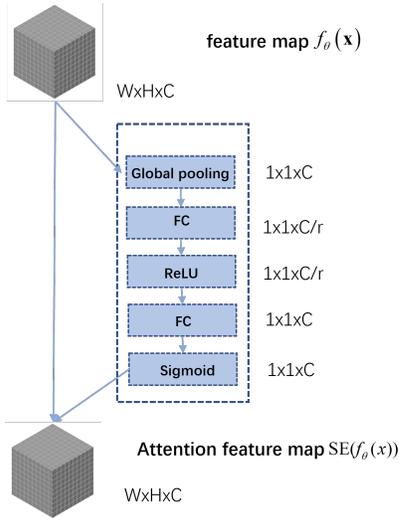
Figure 2: SE Attention

and has strong generalization ability for different data sets. The interdependence between feature channels is explicitly modeled, and the importance of each feature channel is automatically obtained by learning. We take advantage of this to get important information from the feature map that focuses on the distinction between foreground and background, enhance useful features and suppress features that are not useful to the current task. The representation formula of this part is as follows as Eq. (2).

$$AP = SE(f_\theta(x)) \qquad (2)$$

Memory Bank We introduce a memory architecture for few-shot medical image segmentation. In the beginning, we denote the memory bank as M and initial it with an empty queue whose max volume is n for each class, where n is a hyperparameter, and we assign it with 5000. When the number of prototypes in M is less than n, the new prototypes will be inserted into it directly. After filling up, we will sort the queue by their contributions for segmentation and replace the lowest k prototypes with new high contribution prototypes for each class. To be specific, we will first compute the cosine similarity by $sim(\cdot, \cdot)$ between $p_k\left(c^L\right)$ and the prototypes stored in M with the same label. Then, sum the similarity for each prototype in M as their scores to be used to sort later. We can formulate it as Eq. (3) shows.

$$S_i = \sum_{i=0}^{n} M_i\left(c^j\right) \times \mathcal{P} \qquad (3)$$

Similarity-based Segmentation Inspired by (Ouyang et al. 2020), a similarity-based classifier $sim(\cdot, \cdot)$ is introduced, which uses the local image information in $\mathcal{P}$ to predict the query. This is achieved by first matching each prototype to the corresponding local region in the query, and then fusing the local similarities together. Specifically, $sim(\cdot, \cdot)$ first takes query feature map $f_\theta(\mathbf{x}^q)$ and prototype ensemble $\mathcal{P} = \{p_k(c^j)\}$ as input to compute local similarity maps

$S_k(c^j)$ between $f_\theta(\mathbf{x}^q)$ and all $p_k(c^j)'$ s respectivelyy. Each entry $S_k(c^j)_{(h,w)}$ at the spatial location (h, w) corresponding to $f_\theta(\mathbf{x}^q)$ is given by Eq.(4).

$$S_k(c^j)_{(h,w)} = \alpha p_k(c^j) \odot f_\theta(\mathbf{x}^q)(h, w) \qquad (4)$$

where $\odot$ denotes cosine similarly, a multiplier, which helps gradients to backpropagate in training. Then, to obtain similarity maps (unnormalized) with respect to each class c^j as a whole, local similarity maps $S_k(c^j)$ are fused into class-wise similarities $S'(c^j)$, follow as Eq.(7).

$$S'(c^j)_{(h,w)} = \sum_{k} S_k(c^j)_{(h,w)} softmax_k[S_k(c^j)(h, w)]$$
$$(5)$$

Finally, to get the final dense prediction, the class similarity is finally normalized to probability, as Eq.(6) shows.

$$\hat{\mathbf{y}}^q_{(h,w)} = softmax_j[S'(c^j)_{(h,w)}] \qquad (6)$$

Training Based on Triplet Loss

In previous works, the model only focused on the information in one support image at one episode, which actually does not adequately exploit the information in the whole data set. In this part, we elaborate on how to apply triplet loss on a few-shot medical image segmentation task. Intuitively, There is high similarity between the same organs in different people. Take $\{\mathcal{P}_a, \mathcal{P}_p, \mathcal{P}_n\}$ as inputs to loss function, where the subscript a, p and n represent anchor, positive and negative respectively. At train time, we use three loss functions:1. Cross Entropy loss; 2. Prototype Alignment Regularization loss; 3. Triplet loss to optimize our model. $\mathcal{L}_{ce}$ and $\mathcal{L}_{PAR}$ have achieved a good performance in classical few-shot tasks. Triplet loss could be formulated as Eq.(7).

$$\mathcal{L}_{tri} = \max(d(a, p) - d(a, n) + margin, 0) \qquad (7)$$

Where $d(\cdot, \cdot)$ denotes the cosine distance between two parameters.

Experiments

Datasets To demonstrate the efficiency of our proposed method, we performed evaluations under Abd-MRI, which is from ISBI2019 Combined Healthy Abdominal Organ Segmentation Challenge (Task 5). It contains 20 3D T2-SPIR MRI scans.

To unify experiment settings, all images are reformated as 2D axial slices and resized to 256×256 pixels. Each 2D slice is repeated three times in channel dimension to fit into the network.

Evaluation To measure the overlapping between prediction and ground truth, we employ Dice score (0-100, 0:mismatch; 100: perfectly match), which is commonly used in medical image segmentation research. To evaluate the generalization ability to unseen testing classes, beyond the standard few-shot segmentation experiment setting for medical images established by (setting 1), where testing class might appear as background in training data, we introduce setting 2. In setting 2, we force testing classes (even unlabeled) to
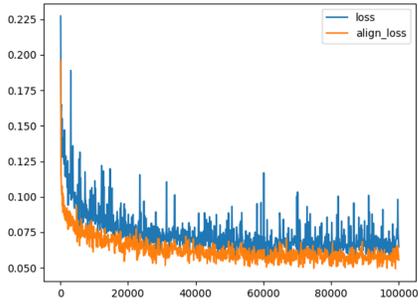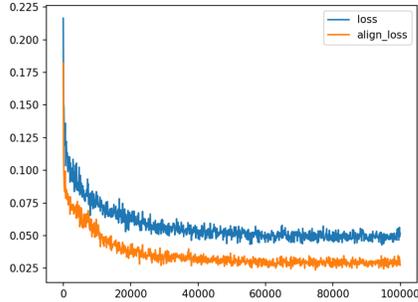
Figure 3: Loss without triplet loss

Figure 4: Loss with triplet loss

be completely unseen by removing any image that contains a testing class, from the training dataset.

Labels are therefore partitioned differently according to the settings and types of supervision. In setting 1, when training with SSL, no label partitioning is required for training. When training with annotated images, each time we take one class for testing and the rest for training. In setting 2, as (spleen, liver), or (left/right kidney) usually appear together in a 2D slice respectively, we group them into the upper abdomen and lower abdomen groups separately. In each experiment, all slices containing the testing group will be removed from the training data.

To simulate the scarcity of labeled data in clinical practice, all our experiments in this section are performed under the 1-way, 1-shot setting.

Implementation Details The implementation is based on the Pytorch (v1.10.2+cu113) implementation of SSL-ALPNet(). The encoder network used is a ResNet-101 pretrained on MS-COCO. Following ALPNet, we optimize the loss using stochastic gradient descent with momentum 0.9, a learning rate of 1e-3 with decay rate of 00.98 per 1k epochs, and a weight decay of 5e-4 over 50k iterations. The self-supervised training takes 6h on a single Nvidia RTX 2080Ti GPU, consuming 4.8GBs of Memory.

Quantitative and Qualitative Results Table I shows the comparisons of our method with SSL-ALP Net, one of the state-of-the-art methods of few-shot medical image segmentation. By using memory architecture, our proposed memory-based SSL-ALP Net occurs overfitting problem. As shown in Fig.2, the results on organs with various shapes, sizes, and intensities have been shown. There is some mis-
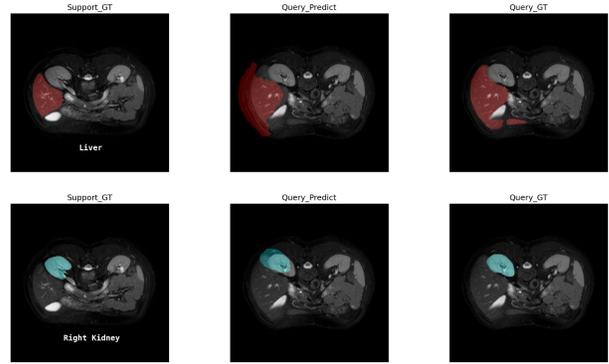

Figure 5: Qualitative results of our method on Abdominal MRI. The proposed method achieves not such desirable segmentation results which are close to ground truth.

matching between foreground and background. As shown in Fig.3, The plot shows the loss values during training. It is obvious that model learning with triplet loss has more stable loss values, which means triplet loss has guided model learning.

| Method | Abdominal-MRI | | | | Mean |
| | Lower | | Upper | | |
	LK	RK	Spleen	Liver	
SSL-ALP(paper)	73.63	78.39	67.02	73.05	73.02
SSL-ALP(rework)	80.95	82.91	69.26	72.12	76.31
memory-based SSL-ALP	74.34	77.89	68.13	71.96	73.12

Table 1: Experimental results on abdominal images.

Conclusion

In this paper, we take a close look at the few-shot image segmentation problem using triplet loss. We also utilize an attention mechanism to further improve the ability of the complementation The proposed method has little effect on model learning. However, we can observe that the model uses Triplet loss to update only the deep layers of the model in the incremental sessions to further mitigate the catastrophic forgetting and overfitting problems. according to the loss values. Evaluation of our model on other medical image datasets in the real world and more benchmark datasets will be one of our future works.

References

Bienenstock, E.; Geman, S.; and Potter, D. 1996. Compositionality, MDL Priors, and Object Recognition. *neural information processing systems*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *neural information processing systems*.

Deng, F.; Feng, H.; Liang, M.; Feng, Q.; Yi, N.; Yang, Y.; Gao, Y.; Chen, J.; and Lam, T. L. 2022. Abnormal Occupancy Grid Map Recognition using Attention Network. In *2022 International Conference on Robotics and Automation (ICRA)*, 8666–8672. IEEE.

Gutiérrez-Viedma, Á.; Parejo-Carbonell, B.; Romeral-Jiménez, M.; Sanz-Graciani, I.; Serrano-García, I.; Cuadrado, M.-L.; and García-Morales, I. 2021. Therapy delay in status epilepticus extends its duration and worsens its prognosis. *Acta Neurologica Scandinavica*, 143(3): 281–289.

Hansen, S.; Gautam, S.; Jenssen, R.; and Kampffmeyer, M. 2022. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78: 102385.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2022. Momentum Contrast for Unsupervised Visual Representation Learning. *computer vision and pattern recognition.*

Hu, J.; Shen, L.; and Sun, G. 2018a. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, J.; Shen, L.; and Sun, G. 2018b. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Isensee, F.; Petersen, J.; Kohl, S. A.; Jäger, P. F.; and Maier-Hein, K. H. 2019. nnu-net: Breaking the spell on successful medical image segmentation. *arXiv preprint arXiv:1904.08128*, 1(1-8): 2.

Li, Y.; Li, Y.; He, W.; Shi, W.; Wang, T.; and Li, Y. 2021. SE-OHFM: A surgical phase recognition network with SE attention module. In *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 608–611. IEEE.

Lieb, D.; Lookingbill, A.; and Thrun, S. 2005. Adaptive Road Following using Self-Supervised Learning and Reverse Optical Flow. In *Robotics: science and systems*, volume 1, 273–280.

Liu, B.; Zhu, Y.; Song, K.; and Elgammal, A. 2021. Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. *Learning.*

Liu, Y.; Zhang, X.; Zhang, S.; and He, X. 2020. Part-aware Prototype Network for Few-shot Semantic Segmentation. *european conference on computer vision.*

Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D.; Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; and Terzopoulos, D. 2020. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Mondal, A. K.; Dolz, J.; and Desrosiers, C. 2018. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241.*

Ouyang, C.; Biffi, C.; Chen, C.; Kart, T.; Qiu, H.; and Rueckert, D. 2020. Self-Supervision with Superpixels: Training Few-shot Medical Image Segmentation without Annotation. *european conference on computer vision.*

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Roy, A. G.; Siddiqui, S.; Pölsterl, S.; Navab, N.; and Wachinger, C. 2019. 'Squeeze & Excite' Guided Few-Shot Segmentation of Volumetric Images. *arXiv: Computer Vision and Pattern Recognition.*

Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; and Boots, B. 2017. One-Shot Learning for Semantic Segmentation. *british machine vision conference.*

Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J. M.; Wang, W.; and Webb, R. 2016. Learning from Simulated and Unsupervised Images through Adversarial Training. *computer vision and pattern recognition.*

Siam, M.; Doraiswamy, N.; Oreshkin, B. N.; Yao, H.; and Jagersand, M. 2020. Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings. *arXiv preprint arXiv:2001.09540.*

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Tang, H.; Liu, X.; Sun, S.; Yan, X.; and Xie, X. 2021. Recurrent Mask Refinement for Few-Shot Medical Image Segmentation. *international conference on computer vision.*

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9197–9206.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. *arXiv: Computer Vision and Pattern Recognition.*