A self-guided anomaly detection-inspired few-shot segmentation network

Suaiba Amina Salahuddin[1,*], Stine Hansen[1], Srishti Gautam[1], Michael Kampffmeyer[1] and Robert Jenssen[1]

[1]*Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø NO-9037, Norway*

Abstract

Standard strategies for fully supervised semantic segmentation of medical images require large pixel-level annotated datasets. This makes such methods challenging due to the manual labor required and limits the usability when segmentation is needed for new classes for which data is scarce. Few-shot segmentation (FSS) is a recent and promising direction within the deep learning literature designed to alleviate these challenges. In FSS, the aim is to create segmentation networks with the ability to generalize based on just a few annotated examples, inspired by human learning. A dominant direction in FSS is based on matching representations of the image to be segmented with prototypes acquired from a few annotated examples. A recent method called the ADNet, inspired by anomaly detection only computes one single prototype. This prototype captures the properties of the foreground segment. In this paper, the aim is to investigate whether the ADNet may benefit from more than one prototype to capture foreground properties. We take inspiration from the very recent idea of self-guidance, where an initial prediction of the support image is used to compute two new prototypes, representing the covered region and the missed region. We couple these more fine-grained prototypes with the ADNet framework to form what we refer to as the self-guided ADNet, or SG-ADNet for short. We evaluate the proposed SG-ADNet on a benchmark cardiac MRI data set, achieving competitive overall performance compared to the baseline ADNet, helping reduce over-segmentation errors for some classes.

Keywords

Medical image segmentation, Few-Shot, Self-supervision

1. Introduction

Significant advances have been made toward image classification and semantic segmentation tasks by deep convolutional neural network-driven approaches such as U-Net, V-Net, FCN and 3D U-Net [1, 2, 3, 4]. Standard fully supervised semantic segmentation strategies can be impractical particularly for medical images as they require large pixel-level annotated datasets which are expensive and time-consuming to acquire, needing considerable clinical expertise. Furthermore, once trained, the models suffer from poor generalisability to classes not encountered during training.

Few-shot learning (FSL) is a promising way to address these challenges. Inspired by how humans are able to distinguish a new concept with just a handful of examples, FSL seeks to learn

The 11th Colour and Visual Computing Symposium, September 08-09, 2022, Gjøvik, Norway

*Corresponding author.

suaiba.a.salahuddin@uit.no (S. A. Salahuddin); s.hansen@uit.no (S. Hansen); srishti.gautam@uit.no (S. Gautam); michael.c.kampffmeyer@uit.no (M. Kampffmeyer); robert.jenssen@uit.no (R. Jenssen)

a model that uses only a single or few annotated samples to segment images from previously unseen classes. Initially, FSL was applied for classification tasks. A key effort was presented in [5]. This then layed down a basis for more recent applications on the more challenging task of semantic segmentation. Most existing few-shot segmentation (FSS) techniques adopt so-called prototypical learning [6, 7, 8, 9, 10, 11]. These approaches usually entail a two-branch encoder-decoder architecture (support and query branches). Here, support refers to the set of images with a few annotated images of certain classes which help learn desired segmentation tasks. The query set refers to the set of images to be segmented, composed of one or more of the same classes as the support. Typically, within a standard prototype-based FSS framework, the support branch extracts class-wise prototypes from the support image which then guides the segmentation of the query image in the query branch. Usually, global average pooling (GAP) is used to extract the prototypes and the query image is segmented based on e.g. cosine similarity between query pixels and the prototypes in the embedding space.

For the particular task of medical image segmentation by FSS, [12] recently proposed self-supervised training with supervoxel pseudo-labels to generate support and query sets for the training phase. Supervoxels refer to groups of similar image voxels and were generated offline as proposed in [13]. The idea is to use an unlabeled image slice and one of its random supervoxel segmentations (as the foreground mask) to be the support image-label pair. Then the query image-label pair is constructed by arbitrary transformations performed on the support data. During testing, new classes are segmented using only a few annotated image slices.

A key drawback of the prototypical FSS approaches is the loss of intra-class local information after GAP. In the context of medical images with large, spatially variant background classes containing all classes other than the foreground this is particularly disadvantageous. Several approaches have attempted to address this with additional prototypes learned class-wise. To tackle this issue and boost segmentation accuracy, [12] incorporated adaptive local prototype pooling. With this strategy, local prototypes were evaluated within a local pooling window overlaid over support data. Recently, [14] postulated that the background volume characteristics cannot be sufficiently represented by prototypes evaluated from only a few support image slices. To overcome this limitation they proposed a novel anomaly detection-inspired technique (ADNet) whereby background prototypes are excluded and only one representative prototype is extracted from the more homogeneous foreground class (i.e. an organ). Then anomaly scores are computed between this foreground prototype and each query pixel to evaluate dissimilarity. In this scheme, the segmentation is then performed by thresholding anomaly scores with a learned threshold value. This approach also includes a novel 3D super-voxel based self-supervision scheme to leverage volumetric data from medical images.

However, the foreground may be too complex to be modelled using a single prototype. This was supported by [15], showing that a single extracted support prototype carried insufficient information to obtain accurate query segmentations even in the case of using the same image as both support and query. They argued, that using average pooling operations inevitably lost useful information needed as support for some query pixels. They sought to overcome this using a self-guided module (SGM) which first extracted a prediction for the labelled support image using the original prototype and then primary and auxiliary prototypes were extracted with masked GAP from the covered and uncovered foreground areas respectively. The primary and auxiliary support vectors were then combined to achieve boosted query segmentation

performances.

Motivated by the findings of [14] and [15], in this work we propose a self-guided ADNet, or SG-ADNet, for short, which generates more fine-grained representations of the foreground properties. In our SG-ADNet, we adopted an Adaptive Self-Guided Module, ASGM, which differing from [15], determines primary and auxiliary prototype outputs based upon number of covered and uncovered foreground regions. Our objective is to investigate properties and potential benefits of the proposed SG-ADNet, having different prototypes (primary and auxiliary) focusing on different foreground regions of interest, to address the potential drawbacks of the single foreground-prototype ADNet framework.

In summary, the main contributions of this work are:

1. We propose a novel framework, SG-ADNet, to help address limitations of the single foreground prototype based ADNet. We leverage multiple foreground prototypes generated using a novel self guidance module, ASGM, to better account for foreground regions "missed" by a single prototype.
2. We show that the SG-ADNet achieves competitive results relative to the ADNet baseline whilst diminishing over-segmentation in cardiac segmentation tasks.

In Section 2 we introduce FSS and our proposed SG-ADNet. Section 3 presents the experimental setup and data used. Section 4 presents results for the experiments, highlighting properties of the proposed method. Appendix A gives additional results. Finally, Section 5 concludes the paper.

2. Self-guided anomaly detection-inspired few shot segmentation for medical images

In order to present the SG-ADNet and its context, we will for the benefit of the reader review the FSS problem setting (Section 2.1), give a brief explanation of the self-supervision stage which is important in FSS approaches to medical image segmentation (Section 2.2), for then to briefly give an overview of the ADNet (Section 2.3). This puts us in a position to explain how the principle of self-guidance is leveraged to propose the novel SG-ADNet.

2.1. The FSS problem setting

In FSS, the model is typically trained on an annotated set D_{train} with classes C_{train} and then this trained model is used to make predictions on a different test set, D_{test} with new classes C_{test} for which only a few annotated examples are available. The training and testing are typically performed episodically [16]; usually with an N-way-K-shot scheme whereby there are N different classes to be distinguished with K examples of each. An episode incorporates support set S and a query set Q for a particular class. The support set contains K annotated images for each class N, which serves as input. The query set contains a query image containing one or multiple N classes. The model learns from the information in the support set about the N classes to then segment the query set outputting a predicted query mask with height and width dimensions H and W respectively.

2.2. Supervoxel-based self supervision

Due to the particular characteristics of medical images, [12] and later [14] developed training episodes for FSS by a self-supervised learning approach via supervoxels. In particular, we adopt the 3D supervoxel approach put forward by [14]. The motivation is to use 3D supervoxels, which are a collection of similar voxels from localised areas within the image volume, to sample "pseudo"-labels for semantically uniform image locations to then guide training. Supervoxels are computed offline using a 3D extension of the efficient, unsupervised, graph-based image-segmentation algorithm prescribed in [13].

In training, each episode involves one unlabeled image volume along with one of its randomly sampled supervoxel masks, representative of the foreground, to first yield a 3D binary mask. Next, 2D slices including this supervoxel class are sampled from the image volume to make up the support and query images. As in [14], we also apply random transformations to support/query images and exploit volumetric data across slices with this 3D approach enabling added information to be available compared to related 2D variants.

2.3. Anomaly detection-inspired FSS - ADNet

As mentioned, in the ADNet-approach, the support and query images, leveraging self-supervised learning (supervoxels), are embedded into features F^s and F^q respectively. The focus is on the foreground class c, only, within each episode. GAP is hence applied only for this class of interest. To ensure masking can be performed, the support feature maps are resized to the same dimensions as the masks (i.e. (H,W)).

One original foreground prototype $p \in \mathbb{R}^d$, with d indicating the dimensionality of the embedding space, is extracted. This is done as follows:

$$p = \frac{F^s(x,y) \circ \mathbf{y}^{fg}(x,y)}{\mathbf{y}^{fg}(x,y)}, \tag{1}$$

where $\circ$ represents the Hadamard product and $\mathbf{y}^{fg}(x,y) = \mathbb{1}(\mathbf{y}(x,y) = c)$ represents the binary foreground mask. This support foreground prototype is input to our self-guided module, as discussed later when presenting the SG-ADNet.

Furthermore, in the original ADNet, segmentation is based on the foreground prototype by adopting a threshold-based metric learning scheme. This involves evaluating anomaly scores S per query feature vector using a negative, scaled cosine similarity measured to the foreground prototype p for that episode by:

$$S(x,y) = -\alpha d(F^q(x,y), p). \tag{2}$$

Here, $d(x,y)$ represents the cosine similarity between F^q and p and the scaling factor α is set to 20 as in [14]. Then, using a learned variable T, the anomaly scores are thresholded to yield the foreground mask prediction. A shifted Sigmoid $\sigma(.)$, is applied to perform soft thresholding which made this procedure differentiable by $\hat{\mathbf{y}}^q_{fg}(x,y) = 1 - \sigma(S(x,y) - T)$. As a result, query features with anomaly scores below threshold T receives a foreground probability greater than 0.5. The background mask, $\hat{\mathbf{y}}^q_{bg}(x,y)$ is 1-$\hat{\mathbf{y}}^q_{fg}(x,y)$.

Following this, the predicted foreground and background masks are upsampled to the same dimensions as the images (H, W) and then a binary cross-entropy loss for segmentation is employed:

$$L_{seg} = -\frac{1}{HW}\sum_{x,y}[\mathbf{y}^q_{bg}(x,y)\log(\hat{\mathbf{y}}^q_{bg}(x,y)) + \mathbf{y}^q_{fg}(x,y)\log(\hat{\mathbf{y}}^q_{fg}(x,y))]. \tag{3}$$

As prescribed in earlier approaches [7, 8, 14, 12] a regularization prototype alignment loss is also adopted whereby the roles of the support and query images are exchanged, i.e. the predicted query mask guides segmentation of the support image:

$$L_{reg} = -\frac{1}{HW}\sum_{x,y}[\mathbf{y}^s_{bg}(x,y)\log(\hat{\mathbf{y}}^s_{bg}(x,y)) + \mathbf{y}^s_{fg}(x,y)\log(\hat{\mathbf{y}}^s_{fg}(x,y))]. \tag{4}$$

Another loss term is also added to minimise the threshold T as follows $L_t = \frac{T}{\alpha}$. The total loss is hence evaluated as follows: $L = L_{seg} + L_{reg} + L_t$.

2.4. The proposed SG-ADNet

Whereas the ADNet approach involved extracting one prototype representative of each class, we are interested in extracting multiple foreground-only prototypes, however still excluding the background. In order to achieve this, we were inspired by [15] to design our proposed ASGM, for our framework. Figure 1 depicts the ASGM. The ASGM coupled with the ADNet constitutes the proposed SG-ADNet. An overview of the framework is presented in Figure 2.

Training of the proposed strategy is conducted in an end-to-end manner. The first steps involve feature extraction (encoding) with a ResNet-101 backbone. The backbone feature extractor with shared weights is used to embed support and query data into deep feature maps. Then metric based learning is used to perform segmentation in the embedding space. The goal is to best utilise the support information. In our approach, first masked GAP is performed over all support foreground pixels to generate initial support prototypes. These prototypes are then input to our ASGM along with the original support masks. The ASGM produces two new prototypes, primary and auxiliary, based on the true positive and false negative pixels of the predicted support masks respectively. The primary prototype preserves the main support data and gathers the True Positive (TP) predictions. The auxiliary prototypes collect all the 'lost' key information not accounted for by the primary prototypes, (the False Negative (FN) predictions), which could not be predicted using the initial support prototype vector. Thus, by aggregating both the primary and auxiliary prototypes we aim to leverage more comprehensive information that could be useful in segmenting the query images.

After the incorporation of our proposed ASGM, a second threshold and a corresponding loss term L_{t2} are added to the overall loss function. Also, as we now have two query predictions, one from the original and one from the self-guided prototype we have two segmentation losses, L_{seg1} and L_{seg2} respectively. Therefore, the new, complete loss term obtained is as follows: $L = L_{seg1} + L_{seg2} + L_{reg} + L_{t1} + L_{t2}$. Here, the L_{t1} and L_{t2} refer to threshold losses corresponding to learned thresholds T_1 and T_2 for thresholding the anomaly scores of the original and the new, self-guided prototypes respectively. Something also notable is that after

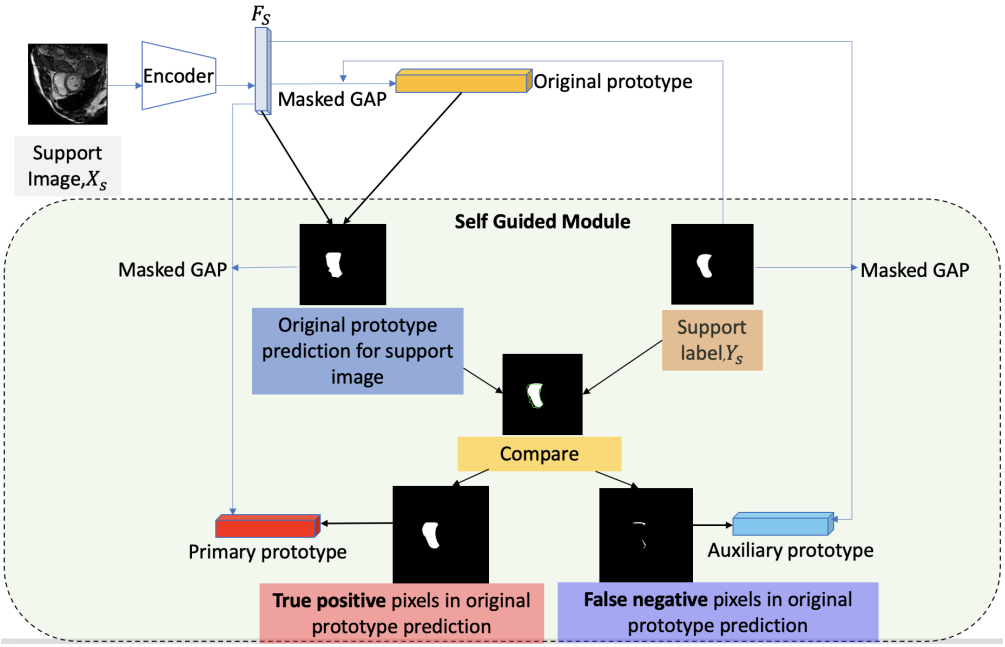

Figure 1: Illustration of the ASGM. The support feature map F_s, the support foreground label Y_s and the original prototype extracted from the support are inputs and produce two new prototypes: primary and auxiliary respectively. The new prototypes are produced by comparing the ground truth and original prototype predictions. The original prototype predictions in the image have been delineated in green and overlaid on the ground truth labels when demonstrating comparison. Here, the primary prototype encodes the true positive regions predicted by the original prototype. The auxiliary prototype encodes the false negative regions predicted by the original prototype.

the ASGM, for the term L_{reg}, we can utilise two predicted query masks (from the original single prototype and the new, self-guided prototype approaches respectively) to compute prototypes for support image segmentation. Note, we only consider the "new" prediction from the combined prototypes and not the original prototype prediction for evaluation.

We investigated different approaches to aggregate the information from the two prototypes (primary and auxiliary). This is represented in the purple box on Figure 2. For our proposed framework, we achieved the best segmentation performance with a weighted sum of the two (primary and auxiliary) prototypes. With this approach, within the ASGM it was determined whether the sums of all evaluated True Positive (TP) and False Negative (FN) pixels exceeded a selected threshold value, τ. Based upon this, four different scenarios were possible with four different ASGM outputs, this is summarised in Figure 3.

The ASGM outputs were determined for each of the four cases as follows:

1. Case 1: both sums of all TP and FN pixels are below τ. Output: original prototype as in [14].
2. Case 2: sum of TP pixels is above τ but the sum of FN pixels is not. Output: primary prototype only.
3. Case 3: sum of FN pixels is above τ but the sum of TP pixels is not. Output: auxiliary

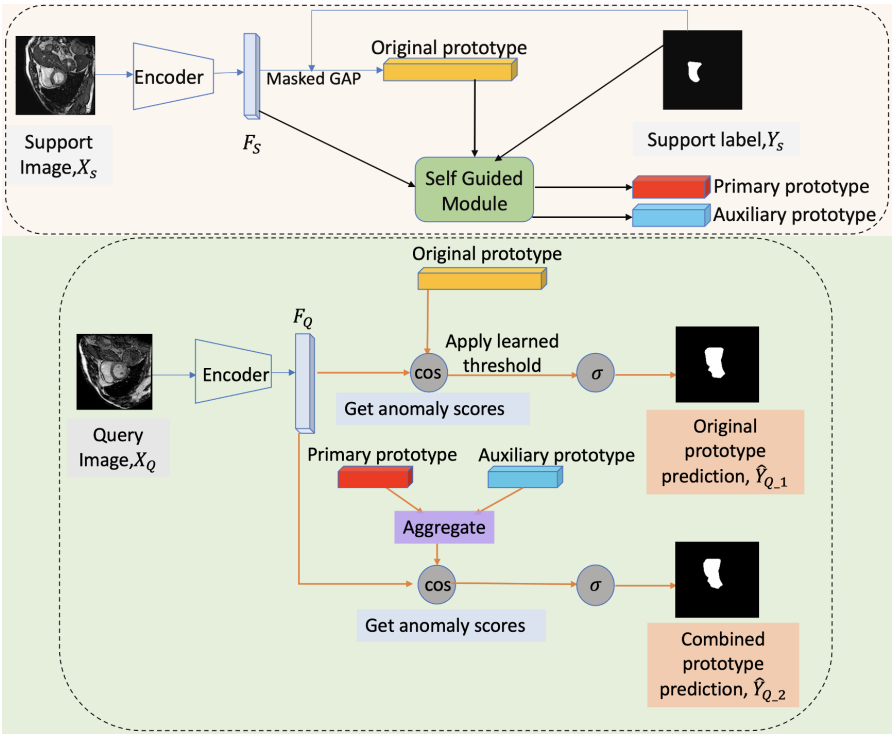

Figure 2: The proposed self-guided, multi prototype FSS framework with anomaly detection, SG-ADNet. In stage I, the prototype extraction and in stage II obtaining the prediction are depicted. Note, that we only consider the "new" prediction from the combined prototypes and not the original prototype prediction for evaluation.

prototype only.

4. Case 4: sums of both TP and FN pixels are above τ. Output: weighted sum of primary and auxiliary prototypes. Here weighting for primary and auxiliary prototypes were: $w_1 = \frac{TP}{TP+FN}$ and $w_2 = \frac{FN}{FN+TP}$ respectively.

We also investigated different settings where certain loss terms were excluded from the total loss to determine the optimal configuration for our SG-ADNet. We found the optimal setting to be as follows: $L = L_{seg1} + L_{seg2} + L_{reg} + L_{t2}$. Here, the L_{t1} term is excluded as we do not want to constrain the learning of the model too much. In addition, the threshold T_1 and the corresponding original prototype are not used in our final predictions. The term L_{reg} is evaluated using sum of two losses obtained with the original and combined query predicted masks guiding support image segmentations respectively.

3. Experiments

Our framework is evaluated on the MS-CMRSeg (bSSFP fold) dataset from the MICCAI 2019 Multi-sequence Cardiac MRI Segmentation Challenge [17]. This dataset contains 35 clinical 3D cardiac MRI scans with labels for the 3 classes Right-Ventricle (RV), Left-Ventricle blood-pool

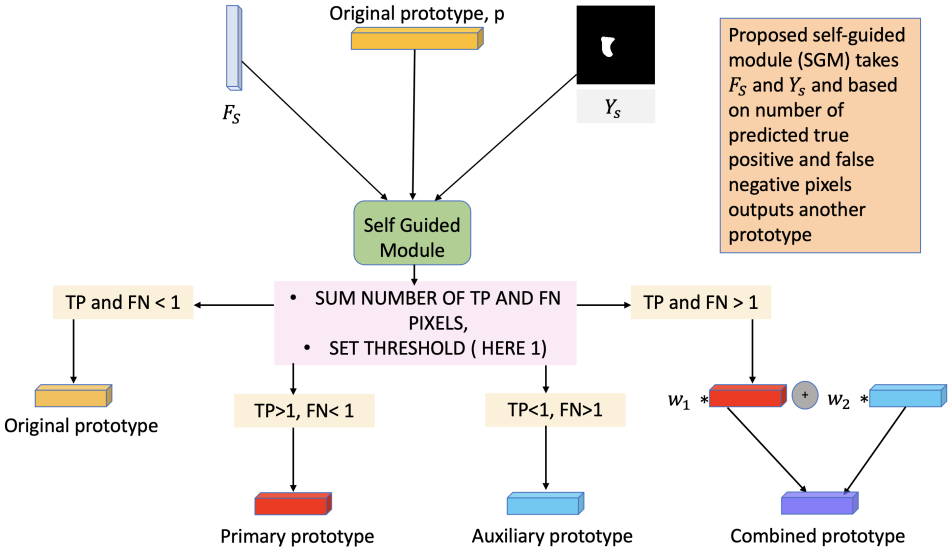

Figure 3: Overview of proposed ASGM's case-based prototype outputs.

(LV-BP) and Left-Ventricle myocardium (LV-MYO). This is a much-used benchmark dataset in FSS for medical images.

For all experiments conducted, self-supervised training is performed and evaluation is done with a 5-fold cross-validation. Per fold, support images are sampled from one subject and the rest are selected as query images.

To account for the stochasticity in the model and optimization, each fold is repeated thrice. Therefore, per fold, the baseline model is repeated thrice and then each baseline model is used to train three runs of SG-ADNet respectively.

Some strategies other than a weighted sum to aggregate the information from the primary and auxiliary prototypes included: stacking the prototypes and similar to [15], concatenating the anomaly scores of the primary and auxiliary prototypes with the query features which was predicted upon using convolutional layers. However, as mentioned above, a weighted sum of the two (primary and auxiliary) prototypes is a good choice, in our experience. Investigations were also performed with and without pre-training of the model with the single prototype method (ADNet) as prescribed in [14], considered to be a baseline method for this work. The key parameters affected by pre-training included the learning rate and model threshold values, the threshold value learned with the single prototype (T) was used to initialise both thresholds (T1 and T2) in the multi-prototype approach.

Note, additionally, motivated by successful utilisation of intermediate layer outputs in [6], we explored adding outputs of different layers of the ResNet, i.e. final layer output only, the output of layer 4 (penultimate layer) and a combination of layers 3 and 4 only. Ultimately, the standard setting with final layer output only was adopted.

3.1. Implementation details

In the proposed SG-ADNet approach, a weighted sum of self-guided prototypes is used, τ is set to 1 to avoid prototypes computed from only one pixel, L_{reg} is determined using both original and new predictions, and L_{t1} is excluded.

The implementation of the proposed framework is based on the PyTorch implementation of the ADNet [14]. The feature extractor backbone architecture chosen is ResNet-101 pre-trained on MS-COCO [18]. The proposed approach is pre-trained with a baseline model as in ADNet with a single prototype. We use a stochastic gradient descent optimiser with a momentum of 0.9. The proposed model's optimiser is initialised with the learning rate obtained at the end of the pre-training baseline model, for 50 epochs.

3.2. Evaluation metrics

We use two standard evaluation metrics, mean dice scores and Intersection over Union (IoU) scores as per prior approaches [7, 12, 8]. For two segmentation masks A and B, the dice score is as follows: $Dice(A, B) = \frac{2||A \cap B||}{||A||+||B||}$. For two segmentation masks A and B, the IoU score can be expressed as follows: $IoU(A, B) = \frac{||A \cap B||}{||A \cup B||}$. Both dice and IoU scores range from 0 to 1 with 0 indicating no and 1 indicating full overlap between A and B. The two metrics are comparable and both are commonly used for evaluating Few-Shot approaches.

4. Results

For quantitative results, we first report the mean dice and IoU scores obtained over 3 runs performed for each of the 5 folds of MS-CMRSeg data. These are compared to the implementation of the baseline approach in [14], with a single prototype. The results are summarized in Tables 1 and 2.

From Tables 1 and 2, we observe that the proposed approach achieves somewhat higher mean dice and IoU scores compared to the baseline approach. For the classes LV-BP and RV, the results with SG-ADNet show an improvement over the ADNet. However, for the LV-MYO structure the ADNet performs better. Thus, it appears that the incorporation of ASGM improves performance for 2 classes for this dataset, namely RV and LV-BP, at the expense of performance for LV-MYO. However, the SG-ADNet provides an increase in the overall mean dice and IoU results over the three classes.

Since the ADNet is the state-of-the-art method in the literature for this benchmark dataset, we consider this to be promising.

The results may indicate that for classes RV and LV-BP, the foreground representation benefits from a more fine grained approach when it comes to prototypes by leveraging self-guidance, and that the opposite effect observed on LV-MYO is not strong enough to hinder SG-ADNet from performing well on the overall mean as compared to ADNet in this case.

For the benefit of the reader, we include qualitative results from one representative example image slice from the cardiac MRI dataset. This is illustrated in Figure 4. The image is visualised at the mid-slice level. Each prediction made, is visualised overlaid on the Ground Truth (GT) labels. The predictions are outlined with a green border and filled in with a red colour. If the

Table 1
Dice scores for proposed and baseline approaches averaged over 3 runs for each of the 5 folds. Highest dice scores are reported in bold lettering.

Method	LV-MYO	LV-BP	RV	Mean
ADNet	**0.595**	0.832	0.645	0.691
SG-ADNet	0.585	**0.840**	**0.697**	**0.707**

Table 2
IoU for proposed and baseline approaches averaged over 3 runs for each of the 5 folds. Highest IoU scores are reported in bold lettering.

Method	LV-MYO	LV-BP	RV	Mean
ADNet	**0.426**	0.717	0.485	0.543
SG-ADNet	0.417	**0.728**	**0.516**	**0.554**

GT and prediction intersects, the colour observed is a pale red and if the prediction exceeds the GT (i.e. over-segmentation occurred) a dark red colour is seen. From Figure 4, for all 3 regions of interest, the proposed model's predictions resembles the GT labels more than the predictions made with a single prototype. The proposed model predictions appears to rectify some over-segmentation in the predictions made with a single prototype. The evaluated dice scores for RV, LV-BP and LV-MYO for this example image are 0.714, 0.883 and 0.631 respectively.

For completeness, we provide further details and results of the investigations. These are presented in the Appendix with the aim to: i) Compare the different strategies we explore on how to aggregate the data from the primary and auxiliary prototypes, and ii) The effect of design choices such as pre-training, different final loss terms and the choice of hyperparameters in our proposed model.

5. Conclusion

In this work, we have proposed SG-ADNet, a novel self-guided anomaly detection-inspired few-shot segmentation framework utilising multiple foreground prototypes. We have particularly investigated the proposed approach for cardiac MRI segmentation. Qualitative results showed the effectiveness of the SG-ADNet in correcting over-segmentation for all structures segmented over the baseline.

Notably, the quantitative results showed improvements in dice and IoU scores for two classes, RV and LV-BP, over the baseline approach. However, this improvement was at the expense of a reduced score for the LV-MYO class, relative to the baseline. Further research should be conducted to better understand these results and the effects of the self-guidance on the cardiac segmentation performance (i.e why the results improved for only RV and LV-BP while they declined for LV-MYO). In future work, we aim to explore additional medical applications and datasets to assess SG-ADNet's full potential and will analyse how it generalises to other imaging modalities.

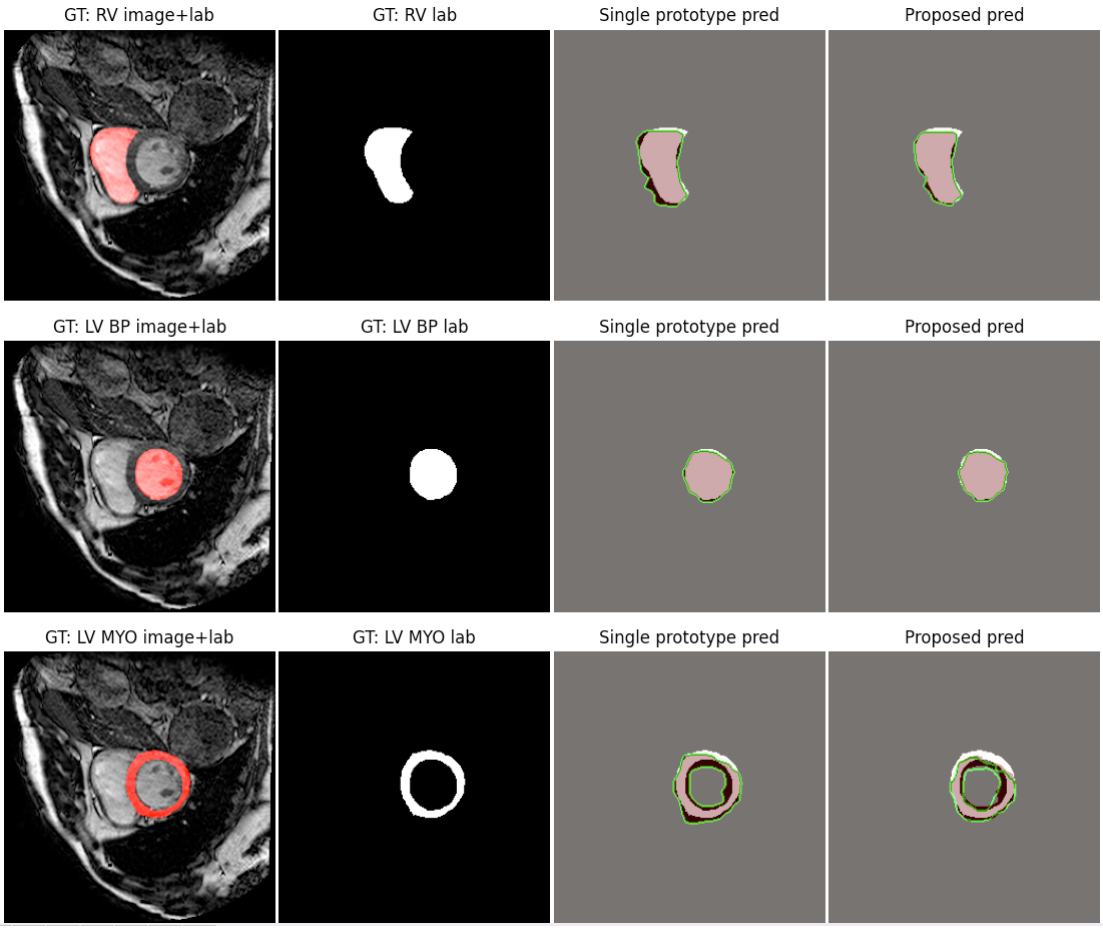

Figure 4: Qualitative results on an example image slice from the cardiac MRI data. Top to bottom the regions of interest are: RV, LV-BP and LV-MYO. From left to right within each row, the visualisations are: i) The original image with the GT label overlaid. The GT label here is highlighted with a red colour, ii) The GT label, iii) The prediction obtained with a single prototype overlaid on the GT label. The prediction here is outlined with a green border and filled in with a red colour & iv) the prediction of the proposed approach with multiple prototypes overlaid on the GT label. Again, the prediction here is outlined with a green border and filled in with a red colour.

Acknowledgments

This work was supported by The Research Council of Norway (RCN), through its Centre for Research-based Innovation funding scheme [grant number 309439] and Consortium Partners; RCN FRIPRO [grant number 315029]; RCN IKTPLUSS [grant number 303514]; and the UiT Thematic Initiative.

References

[1] O. Ronneberger, P. Fischer, 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. Milletari, N. Navab, 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.

[3] W. Sun, R. Wang, Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm, IEEE Geoscience and Remote Sensing Letters 15 (2018) 474–478.

[4] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer, 2016, pp. 424–432.

[5] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Advances in Neural Information Processing Systems 2017-December (2017) 4078–4088.

[6] C. Zhang, G. Lin, F. Liu, R. Yao, C. Shen, Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[7] K. Wang, J. H. Liew, Y. Zou, D. Zhou, J. Feng, Panet: Few-shot image semantic segmentation with prototype alignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[8] Y. Liu, X. Zhang, S. Zhang, X. He, Part-aware prototype network for few-shot semantic segmentation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12354 LNCS (2020) 142–158. doi:10.1007/978-3-030-58545-7_9.

[9] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, J. Kim, Adaptive prototype learning and allocation for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8334–8343.

[10] Q. Yu, K. Dang, N. Tajbakhsh, D. Terzopoulos, X. Ding, A location-sensitive local prototype network for few-shot medical image segmentation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 262–266.

[11] F. Cermelli, M. Mancini, Y. Xian, Z. Akata, B. Caputo, A few guidelines for incremental few-shot segmentation, arXiv e-prints (2020) arXiv–2012.

[12] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, D. Rueckert, Self-supervision with superpixels: Training few-shot medical image segmentation without annotation, volume 12374 LNCS, Springer Science and Business Media Deutschland GmbH, 2020, pp. 762–780. doi:10.1007/978-3-030-58526-6_45.

[13] P. F. Felzenszwalb, D. P. Huttenlocher, Efficient graph-based image segmentation, International journal of computer vision 59 (2004) 167–181.

[14] S. Hansen, S. Gautam, R. Jenssen, M. Kampffmeyer, Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels, Medical Image Analysis 78 (2022) 102385.

[15] B. Zhang, J. Xiao, T. Qin, Self-guided and cross-guided learning for few-shot segmentation,

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021 (2021) 8312–8321.
[16] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, Advances in neural information processing systems 29 (2016).
[17] X. Zhuang, Multivariate mixture model for cardiac segmentation from multi-sequence mri, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 581–588.
[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

A. Appendix

The details of experiments conducted, to determine the optimal strategy for aggregating data from the prototypes produced by our ASGM, are presented.

First, the relative performances of the methods we considered for aggregating information from the new primary and auxiliary prototypes obtained with our ASGM are summarised in Table 3. The methods compared include: i) stacking the prototypes with $\tau = 1$, ii) concatenating the anomaly scores the query predictions achieved with each prototype (primary and auxiliary), and passing through two 3×3 convolutional layers and iii) the weighted sum of prototypes as in our proposed approach. The reported metrics are dice scores for fold 1 of the MS-CMRSeg cardiac MRI data. Our proposed strategy of using weighted sum was chosen as it had an overall good performance across all three classes. The proposed weighted sum approach had the best performance, amongst the methods compared, for the LV-BP and RV classes and had the second-best performance for LV-MYO compared to the baseline.

Next, Table 4 summarises mean dice scores over 5 folds obtained with different combinations of the final loss term and thresholds applied. In each investigated approach, the weighted sum of primary and auxiliary prototypes as described previously was adopted. Three settings of L_{reg} were explored: using original prediction only, using new prototype only and using both predictions. Effect of whether or not to: i) initialise learning rate with pre-training, ii) include L_{seg1}, iii) include L_{t1}, iv) include L_{t2} and v) keeping t1 fixed to pre-trained value were investigated.

Comparing the dice score results reported, the most appropriate approach determined was: the initialisation of learning rate with pre-training, L_{reg} evaluated with both original and new,

Table 3
Comparison of investigated strategies for aggregating data from primary and auxiliary prototypes for fold 1. Highest dice scores in bold.

Aggregation approach	LV-MYO	LV-BP	RV
Stacked prototypes	0.527	0.760	0.641
Concatenation of anomaly scores with predictions	0.530	0.832	0.602
Weighted sum of prototypes (proposed)	**0.595**	**0.861**	**0.706**

Table 4
Investigation of different final loss terms for pre-trained, weighted sum prototype approach to determine optimal configuration. The evaluation metric reported here is the dice score. Highest dice scores are reported in bold lettering. The selected approach for the proposed scheme is shown in the last row.

$\mathbf{L_{seg1}}$	T_1 **fixed**	$\mathbf{L_{t1}}$	$\mathbf{L_{t2}}$	**LV-BP**	**LV-MYO**	**RV**
✓	✓	✓	✓	0.600	0.857	0.703
✓	×	×	✓	0.595	**0.861**	0.706
✓	×	✓	✓	0.605	0.852	0.698
✓	×	×	×	**0.612**	0.841	0.682
×	×	✓	✓	0.583	0.860	**0.710**
✓	×	×	✓	0.595	**0.861**	0.706

self-guided predictions, the inclusion of L_{seg1}, and L_{t2}, excluding L_{t1} and not fixing t1. This selected approach is shown in the last row of Table 4.