Few-shot Semantic Segmentation of Wireless Capsule Endoscopy Images

Kevin Raj. P[1] and Chandra Sekhar Seelamantula[1]

Department of Electrical Engineering, Indian Institute of Science, India
{kevinraj,css}@iisc.ac.in

Abstract. The main challenge in few-shot learning is the translation of latent representations learned from the support image using a feature extractor to the query image. To achieve it, we propose an end-to-end Bayesian learning framework for few-shot learning, which utilizes the information from both the support and query image. The features extracted from the support (prior) and query image (posterior) are modeled as a multivariate Gaussian Mixture Model (GMM) using an auto-encoder coupled with a shallow convolutional neural network. Afterward, the support and query prototypes are sampled from the learned GMM distribution and fused with the extracted query features to estimate the final segmentation map. The joint optimization of the feature maps and the GMM parameters results in rich feature extraction and robust distribution estimation of the input samples. It also alleviates the network from finding the local optima, strengthening the overall stability of the network. The proposed technique is extensively validated on two publicly available wireless capsule endoscopy datasets, KID-1 consisting of 77 images of 9 different abnormalities & KID-2 consisting of 593 images of 4 different abnormalities proving the efficacy of our technique. The code will be released at https://github.com/kevinYitshak/wce

Keywords: Few-shot · Segmentation · Gaussian Mixture Models · Wireless Capsule Endoscopy.

1 Introduction

Assessment of gastrointestinal (GI) tract was tiring and painful for the patients until the introduction of wireless capsule endoscopy (WCE) by Iddan et al. [10]. WCE considerably reduced the difficulty in the visualization of the small bowel region, where the possibility of the occurrence of various diseases is relatively high. Patients are required to ingest a WCE capsule for 7-8 hours, resulting in 50,000 frames [1]. Analyzing 50,000 frames per patient for gastroenterologists can be demanding, which led to many automated approaches.

1.1 Literature Review

Wireless Capsule Endoscopy. Early works such as [6,17] made use of handcrafted features like histogram, mean, variance, of different color-spaces, texture information, and finally used hidden Markov models, support vector machines for classification and thresholding techniques for the segmentation task. Regardless, manual feature selection does not always produce desirable results. Methods like [21,11] use deep convolutional neural networks, leading to vast performance gains for the classification and localization of abnormalities. But for the segmentation task, most of the previous works are solely focused on specific diseases like bleeding [15,7], ulcer [27,3], or polypoid [13,24,5]. In our prior works [18,22] we proposed a four-channel U-Net consisting of RGB-alpha color space and patch-based feature extraction using a convolutional neural network using the standard training-testing procedure. The main challenges of the previous work predominantly lack the generalization capability. Also, due to fewer samples per abnormality, the network can lead to over-fitting as the image acquisition and expert annotations are demanding. To alleviate the challenges faced by the traditional techniques, few shot techniques gained more attention recently.

Few-shot techniques. In general, prototype learning techniques like [23,25,16], meta-learning [12,19] and data generation based approaches [29,8] are the most common few-shot methodologies. Our proposed work closely aligns with two recent works: Zhang et al. [28] and Yang et al. [26]. The key difference compared to [28] is instead of class-wise distribution estimation, we used image-wise distribution as medical images can quite vary within the same classes and also exploited the query image distribution for better translation of latent representations from support to query image. Compared to [26], end-to-end learning framework is proposed resulting in better feature extraction further leading to better distribution estimation and performance.

1.2 Our Contribution

Our contribution consists of three folds: 1) We propose a novel Bayesian network for a few-shot segmentation task by utilizing the semantics of both the support and query image. 2) End-end few-shot pipeline is proposed for better feature extraction and estimation of GMM distribution. 3) Compared to previous works in WCE segmentation, we provide an extensive validation on two different public WCE datasets, consisting a total of 11 abnormalities.

2 Proposed Method

2.1 Problem Formulation

Let's consider a dataset $D(X, Y)$ consisting of C classes, with each class containing images X and its corresponding masks Y. In few-shot paradigm, D is separated into two non-overlapping sets namely, *support set* $S = \{(x_i^s, y_i^s)\}_i^{C,K} \subset X, Y$ of K randomly sampled images and masks of class C and *query set* $Q = \{(x_i^q, y_i^q)\}_i^{C,N} \subset X, Y$ of other N images and masks of the same class C; y_i^q is not used during the training phase. Distinctly, the data $D(X, Y)$ from classes C seen

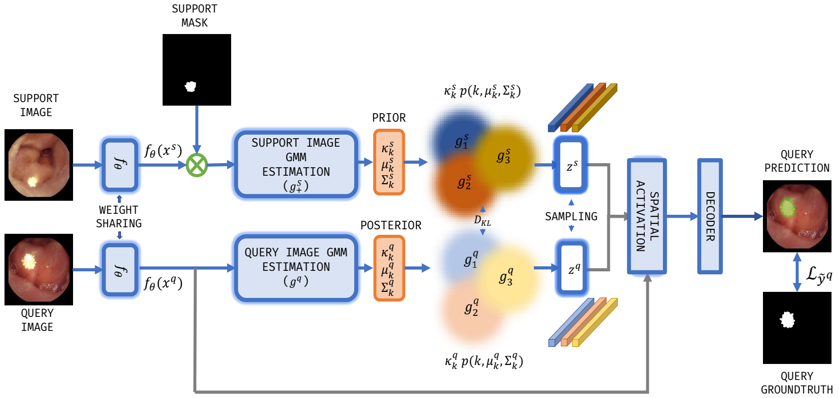

SUPPORT MASK
SUPPORT IMAGE
f_θ
$f_\theta(x^s)$
WEIGHT SHARING
SUPPORT IMAGE GMM ESTIMATION (g_+^s)
PRIOR
κ_k^s μ_k^s Σ_k^s
$\kappa_k^s\, p(k, \mu_k^s, \Sigma_k^s)$
g_1^s
g_2^s
g_3^s
D_{KL}
z^s
SAMPLING
z^q
SPATIAL ACTIVATION
DECODER
QUERY PREDICTION
$\mathcal{L}_{\tilde{y}^q}$
QUERY GROUNDTRUTH
QUERY IMAGE
f_θ
$f_\theta(x^q)$
QUERY IMAGE GMM ESTIMATION (g^q)
POSTERIOR
κ_k^q μ_k^q Σ_k^q
g_1^q
g_2^q
g_3^q
$\kappa_k^q\, p(k, \mu_k^q, \Sigma_k^q)$

Fig. 1. [Color online] Proposed network consists of two branch weight sharing feature extractor f_θ one for support & other for query image(s). Foreground support features $f_\theta(x_+^s)$ are spatially partitioned using the support mask y^s. Further spatially activated query regions using the sampled latent representations from the estimated GMM are passed through the decoder block resulting in the final segmentation map $\tilde{y}^q$.

during the training phase is not seen during the test phase, meaning training and test sample classes are mutually exclusive. Once, the network has learned, its parameters are fixed and not optimized further during the testing phase.

Main information required to predict the segmentation mask ($\tilde{y}^q$) of the proposed technique is the latent features sampled from the gaussian mixtures of both, the support and the query distributions. The final prediction for a query image (x^q) is given as:

$$\tilde{y}^q = \{p(f_\theta(x^q), z_q, z_s | x^q, S)\} \tag{1}$$

where z_q, z_s are the sampled latent features from query and support gaussian mixtures, $f_\theta(x^q)$ is the extracted query features from a feature extractor f_θ and $p(.)$ is the probability of a pixel being the foreground.

2.2 Architecture Description

We have used the two-branch shared weight network (with trainable parameters θ) to extract the support $f_\theta(x^s)$ and query $f_\theta(x^q)$ features. The support features $f_\theta(x^s)$ are spatially partitioned into foreground features $f_\theta(x_+^s)$ with the help of support mask y^s. Upon obtaining the features, $f_\theta(x_+^s)$ and $f_\theta(x^q)$ are clustered by assigning soft probabilities to each feature map by modeling it as a gaussian mixture model denoted as g_+^s and g^q with K gaussian distributions. An auto-encoder with a shallow CNN is employed to estimate the GMM parameters κ, μ, Σ to simultaneously optimize both the auto-encoder (with trainable parameters ϕ^s, ϕ^q) and also the mixture model providing an end-to-end training [30], unlike Expectation-Maximization (EM) algorithm.

The latent representations sampled from g^s_+ are given as z^s. But it is not straightforward in the case of g^q because the sampled latent features might not be of our interest ie., it can either belong to the foreground or the background representation spatially. To learn the foreground query features, we consider query features GMM g^q as a posterior and the support features GMM g^s_+ as a prior. During the training phase, we bring g^q closer to g^s_+ by the minimizing the KL-Divergence D_{KL} and make sure the sampled latent representations from g^q belongs to g^s_+. Our task lies in two-fold: 1) Estimation of posterior g^q and prior g^s_+ distribution and 2) How to make use of the estimated distribution to predict the query mask (y^q) as given in Section 2.3 and 2.4.

2.3 Estimation of Distribution

The extracted features $f_\theta(x^s)$ and $f_\theta(x^q)$ are passed through an auto-encoder to perform dimensionality reduction resulting in lower dimension features. In addition, the error between the reconstructed features are also utilized for GMM parameter estimation [30]. For example, let's consider the support features $f_\theta(x^s)$,

$$f^s_l = \mathbb{E}(f_\theta(x^s); \phi^s_e), \quad \widetilde{f}_\theta(x^s) = \mathbb{D}(f^s_l; \phi^s_d), \quad f^s_r = \mathbb{H}(f_\theta(x^s), \widetilde{f}_\theta(x^s)). \tag{2}$$

where f^s_l is the dimensionality reduced feature and $\widetilde{f}_\theta(x^s)$ is the reconstructed feature of $f_\theta(x^s)$. $\mathbb{E}(.)$, $\mathbb{D}(.)$ denotes encoder and decoder functions of the auto-encoder network. $\mathbb{H}(.)$ denotes reconstruction error functions f^s_r, which can be multi-dimensional. In our case, cosine similarity $(\mathcal{S}_c)$ and euclidean distance $(\mathcal{E}_d)$ functions are considered to measure the error features. Finally, ζ^s is passed to the estimation network for calculating the GMM parameters Fig. 2a.

$$f^s_r = [\mathcal{S}_c(f_\theta(x^s), \widetilde{f}_\theta(x^s)), \mathcal{E}_d(f_\theta(x^s), \widetilde{f}_\theta(x^s))], \quad \zeta^s = [f^s_l, f^s_r]. \tag{3}$$

Modelling a Single Image as Mixture of Gaussian's: Given, the low dimensional features of the support image $\zeta^s \in \mathbb{R}^{d \times W \times H}$ the mixture probability is calculated as follows,

$$\mathbf{P^s} = \mathbb{S}(\zeta^s), \quad \widehat{\lambda} = softmax(\mathbf{P^s}). \tag{4}$$

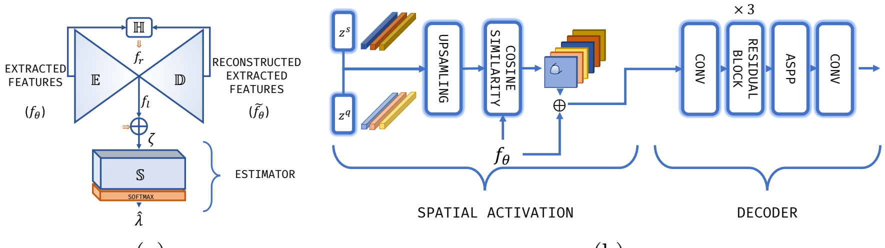
where $\mathbb{S}$ denotes a shallow convolutional neural network with trainable parameter (ϕ^s_s) producing an output $\mathbf{P^s}$. $\widehat{\lambda} \in \mathbb{R}^{K \times W \times H}$ denotes the soft-mixture probability prediction and K is the number of gaussian distributions in a GMM. Using ζ^s and $\widehat{\lambda}$, the parameters are calculated as given below. Note, $N = W \times H$.

$$\kappa^s_k = \sum_{k=1}^{K}\sum_{i=1}^{N} \frac{\widehat{\lambda}_{ik}}{N}, \mu^s_k = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N} \widehat{\lambda}_{ik}\zeta^s_i}{\sum_{k=1}^{K}\sum_{i=1}^{N} \widehat{\lambda}_{ik}}, \Sigma^s_k = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N} \widehat{\lambda}_{ik}(\zeta^s_i - \mu^s_k)(\zeta^s_i - \mu^s_k)^T}{\sum_{k=1}^{K}\sum_{i=1}^{N} \widehat{\lambda}_{ik}}.$$

The mixture of gaussian's is defined for the support features as,

$$g^s_+(\zeta^s|\phi^s_s) = \sum_{k=1}^{K} \kappa^s_k p(\zeta^s|k, \mu^s_k, \Sigma^s_k). \tag{5}$$

$$p(\zeta^s|k, \mu^s_k, \Sigma^s_k) = \frac{1}{(2\pi)^d|\Sigma_k|^{\frac{1}{2}}} exp[-1/2(\zeta^s - \mu^s_k)^T \Sigma^{-1}_k(\zeta^s - \mu^s_k)]. \tag{6}$$



(a) (b)

Fig. 2. [Color online] (a) Gaussian Mixture Model comprising of the encoder $\mathbb{E}(\phi_e)$, decoder $\mathbb{D}(\phi_d)$ and a shallow convolutional network $\mathbb{S}(\phi_s)$, (b) Spatial activation and decoder block.

In a similar fashion, the same set of variables of query features is also estimated κ_k^q, μ_k^q and Σ_k^q. Further, $g^q(\zeta^q|\phi_s^q) = \sum_{k=1}^{K} \kappa_k^q p(\zeta^q|k, \mu_k^q, \Sigma_k^q)$ is calculated. The parameters of the GMM are constrained such that, $\kappa^s, \kappa^q \geq 0$, $\sum_{k=1}^{K}(\kappa_k^s, \kappa_k^q) = 1$ and Σ^s, Σ^q are positive-definite matrix. Note that, $(\kappa_k^s, \kappa_k^q) \in \mathbb{R}^1, (\mu_k^s, \mu_k^q) \in \mathbb{R}^d, (\Sigma_k^s, \Sigma_k^q) \in \mathbb{R}^{d \times d}$. The log-likelihood of the modelled gaussian mixture using the estimation network (ϕ_s^s, ϕ_s^q) and the objective function of the auto-encoder $(\Theta^s = (\phi_e^s, \phi_d^s), \Theta^q = (\phi_e^q, \phi_d^q))$ for support and query features are given as,

$$L^s(\phi_s^s) = -log \sum_{k=1}^{K} \kappa_k^s p(\zeta^s|k, \mu_k^s, \Sigma_k^s), \quad L^q(\phi_s^q) = -log \sum_{k=1}^{K} \kappa_k^q p(\zeta^q|k, \mu_k^q, \Sigma_k^q).$$

$$L_b(\phi_s^s, \phi_s^q) = log \sum_{k=1}^{K} \kappa_k^s p(z^q|k, \mu_k^s, \Sigma_k^s) - D_{KL}(g^q||g_+^s). \tag{7}$$

First term in Eq. 7 constraints the sampled query representations belonging to the foreground region and second term brings the posterior distribution g^q closer to the prior g_+^s by minimising the KL divergence. Total objective function for the GMM estimation is given by,

$$L_{GMM} = J^s(\Theta^s) + J^q(\Theta^q) - \lambda_1 L_b(\phi_s^s, \phi_s^q) + \lambda_2\{L^s(\phi_s^s) + L^q(\phi_s^q)\}. \tag{8}$$

where $J^s(\Theta^s) = L_2(f_\theta(x^s), \widetilde{f}_\theta(x^s))$ and $J^q(\Theta^q) = f_\theta(x^q), \widetilde{f}_\theta(x^q)$. λ_1 and λ_2 being the hyper-parameters set to $1e-2$ and $1e-4$ based on experimentation.

2.4 Output Prediction of Query Image

The latent representations $(z_q, z_s) \in \mathbb{R}^{d \times 1 \times 1}$ are sampled from the estimated support g_+^s and query distribution g^q and interpolated to the same spatial resolution of $f_\theta(x^q)$. In-order to spatially activate the interest regions in the query features, cosine similarity between $[z^q,\ f_\theta(x^q)]$ and $[z^s,\ f_\theta(x^q)]$ are calculated as shown in Fig. 3 and concatenated with the extracted query features $f_\theta(x^q)$. Finally, it is passed through a decoder block Fig. 2b containing Atrous Spatial Pyramid Pooling (ASPP) and residual block resulting in

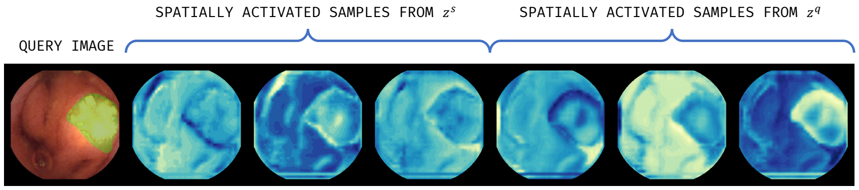

QUERY IMAGE
SPATIALLY ACTIVATED SAMPLES FROM z^s
SPATIALLY ACTIVATED SAMPLES FROM z^q

Fig. 3. [Color online] Spatially activated regions of the query image by the sampled latent representation z^s and z^q. Green overlay on the query image indicates the final prediction.

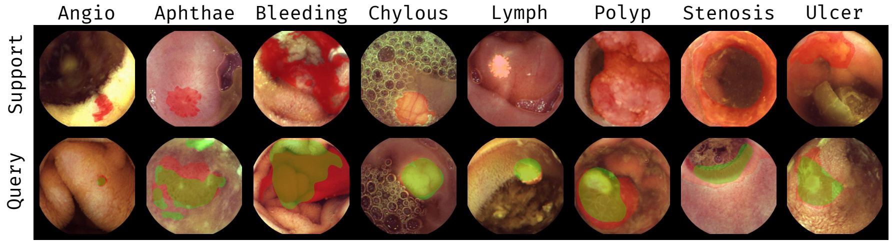


Fig. 4. [Color online] Overlaid segmented results. Green indicates prediction and Red indicates ground-truth.

the final segmentation output $\tilde{y}^q$. The objective function between $\tilde{y}^q$ and y^q is $L_{\tilde{y}^q} = \mathcal{BCE}(y^q, \tilde{y}^q) + \mathcal{IOU}(y^q, \tilde{y}^q)$. The proposed network is optimized by simultaneously minimizing 1) the GMM estimation (L_{GMM}) and 2) Segmentation ($L_{\tilde{y}^q}$) objective function resulting in end-to-end training, given as $L = L_{\tilde{y}^q} + L_{GMM}$.

3 Experiments

3.1 Datasets

The proposed technique is validated on two publicly available datasets named KID-1 [9] and KID-2 [14]. KID-1 consists a total of 77 images containing 9 abnormalities: angioectasias (27), aphthae (5), lymphangiectasia (9), polypoid (6), bleeding (5), chylous (8), stenosis (6), ulcer (9) and, villous oedemas (2) [not considered due to less number of images]. KID-2 consists of 593 images containing four abnormalities: vascular (303), inflammatory (227), polypoid (44) and ampulla-of-vater (19). Polypoid of KID-1 & KID-2 are combined together and considered as KID-1. The images from both the datasets are of spatial dimension 360 × 360 and pixel-wise annotations are used as ground-truth for the few-shot segmentation task. The eight abnormalities in KID-1 are divided into four groups, each group containing two abnormalities. Out of four groups, three groups are used for training and the other one group is used for validation. Abnormalities of KID-2 are completely used for validation.

Table 1. Performance metrics of 6-way 1-shot setting for KID-1 and KID-2 of the proposed technique.

Datasets	Group	Abnormalities	IOU	Dice	Sensitivity	Specificity	Accuracy
KID-1	Group 0	Angioectasia	0.2791	0.3868	0.7867	0.9096	0.907
		Apathe	0.1652	0.2358	0.3446	0.9411	0.8922
	Group 1	Lymphangiectasia	0.5495	0.7061	0.9001	0.9798	0.9743
		Polypoid	0.1627	0.2309	0.2151	0.9635	0.8223
	Group 2	Bleeding	0.3269	0.4767	0.6210	0.7858	0.7478
		Chylous	0.6602	0.7899	0.8130	0.9697	0.942
	Group 3	Stenosis	0.3198	0.4583	0.4133	0.8792	0.7597
		Ulcer	0.4235	0.5322	0.5753	0.9673	0.9435
		Mean	**0.3608**	**0.4770**	**0.5836**	**0.9245**	**0.8736**
KID-2	Group 0	Ampulla-of-vater	0.4593	0.5917	0.8016	0.8997	0.8941
	Group 1	Inflammatory	0.2260	0.3222	0.4673	0.9456	0.8969
	Group 2	Vascular	0.1065	0.1653	0.4290	0.9317	0.9088
		Mean	**0.2639**	**0.3597**	**0.5659**	**0.9256**	**0.8999**

3.2 Implementation Details

The support and query features are obtained using the DeepLab Resnet-50 [4] architecture, pretrained on Imagenet [20]. The proposed model is trained with four pairs of support and query images per batch by optimizing the final objective function L, using Adam optimizer with a learning rate of $3e-4$. During the training phase, the learning rate is reduced by using cosine decay. The model is trained for 200 epochs with an early stopping, based on the validation dice score with a tolerance of 50 epochs. Due to a fewer number of training images, data augmentation is performed using a library called albumentation [2]. The abnormality classes are randomly chosen and the support-query images from the chosen class are also randomly sampled during the training step. Based on experimentation, three multivariate gaussian distributions ($K = 3$) per GMM of dimension ($d = 64$) gave desirable results. For k-shot setting, the GMM parameters for k support images are estimated as per section 2.3 and averaged over the k support images, resulting in the final g^s_+.

3.3 Results

Performance of the proposed technique is evaluated by calculating the standard metrics such as IOU, dice, sensitivity, specificity, and accuracy as given in Table 1. The few-shot paradigm is comparatively new and to our best of our knowledge there is no few-shot segmentation of Wireless Capsule Images. For comparison of the proposed technique we implement a recent few-shot technique FPMMs and it's variant FRPMMs [26] and compare our proposed method as given in Table 2. Our technique achieves an increase of 7.25% for 1-shot and 4.86% for 5-shot in average dice score compared to [26]. It also achieves superior performance for seven abnormalities proving the efficacy of the proposed technique. Moreover, the performance of our proposed 1-shot technique is performing better than the 5-shot of [26]. Comparison of k-shot setting is given in Table 3.2. In

Table 2. Comparison of 6-way 1-shot average dice score for KID-1 and KID-2 dataset abnormality wise with other few-shot techniques.

Datasets	Methods Abnormalities	FPMM	FRPMM	VGMM (ours)
KID-1	Angioectasia	0.3310	0.3596	**0.3868**
	Apathe	**0.3893**	0.1285	0.2358
	Lymphangiectasia	0.2600	0.6360	**0.7061**
	Polypoid	0.2177	**0.4239**	0.2309
	Bleeding	0.5393	**0.5945**	0.4767
	Chylous	0.6163	0.5112	**0.7899**
	Stenosis	0.3981	0.2626	**0.4583**
	Ulcer	0.4842	0.3651	**0.5322**
	Mean	0.4044	0.4101	**0.4770**
KID-2	Ampulla-of-vater	0.4358	0.3771	**0.5917**
	Inflammatory	0.2552	0.2968	**0.3222**
	Vascular	**0.1738**	0.1709	0.1653
	Mean	0.2882	0.2816	**0.3597**

Table 3. Comparison of average dice score of 6-way 1-shot and 6-way 5-shot settings.

Datasets	k-shot	FPMMs	FRPMMs	VGMM(ours)
KID-1	1-shot	0.4044	0.4101	**0.4770**
	5-shot	0.4069	0.4255	**0.4669**
KID-2	1-shot	0.2882	0.2816	**0.3597**
	5-shot	0.2941	0.3073	**0.3631**

contrary, performance gain of the proposed method between 1-shot and 5-shot is quite negligible, which can considered as a task for future works. The final segmentation results of our proposed technique is given in Fig 4.

4 Conclusion

Due to very few data samples, traditional semantic segmentation of WCE can easily lead to over-fitting, and the generalization capability of the network is greatly compromised. In order to overcome these issues, we have proposed a few-shot based network and to the best of our knowledge, there are no previous works related to few-shot semantic segmentation of WCE images. Our work is extensively validated on two public datasets, KID-1 and KID-2 containing 11 abnormalities in total achieving a 7.25% for 1-shot and 4.86% for 5-shot, increase in average dice and also achieves better performance for 7 abnormalities, thus proving the generalization capability and efficacy of the proposed technique.

References

1. Adler, D.G., Gostout, C.J.: Wireless capsule endoscopy. Hospital Physician **39**(5), 14–22 (2003)
2. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information **11**(2), 125 (2020)
3. Charfi, S., El Ansari, M., Balasingham, I.: Computer-aided diagnosis system for ulcer detection in wireless capsule endoscopy images. IET Image Processing **13**(6), 1023–1030 (2019)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
5. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 263–273. Springer (2020)
6. Fu, Y., Zhang, W., Mandal, M., Meng, M.Q.H.: Computer-aided bleeding detection in wce video. IEEE journal of biomedical and health informatics **18**(2), 636–642 (2013)
7. Ghosh, T., Li, L., Chakareski, J.: Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 3034–3038. IEEE (2018)
8. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3018–3027 (2017)
9. Iakovidis, D.K., Koulaouzidis, A.: Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software. Gastrointestinal endoscopy **80**(5), 877–883 (2014)
10. Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless capsule endoscopy. Nature **405**(6785), 417–417 (2000)
11. Jain, S., Seal, A., Ojha, A., Krejcar, O., Bureš, J., Tachecí, I., Yazidi, A.: Detection of abnormality in wireless capsule endoscopy images using fractal features. Computers in Biology and Medicine **127**, 104094 (2020)
12. Jamal, M.A., Qi, G.J.: Task agnostic meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11719–11727 (2019)
13. Jia, X., Xing, X., Yuan, Y., Xing, L., Meng, M.Q.H.: Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition. Proceedings of the IEEE **108**(1), 178–197 (2019)
14. Koulaouzidis, A., Iakovidis, D.K., Yung, D.E., Rondonotti, E., Kopylov, U., Plevris, J.N., Toth, E., Eliakim, A., Johansson, G.W., Marlicz, W., et al.: Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. Endoscopy international open **5**(6), E477 (2017)
15. Li, S., Zhang, J., Ruan, C., Zhang, Y.: Multi-stage attention-unet for wireless capsule endoscopy image bleeding area segmentation. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 818–825. IEEE (2019)
16. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. In: European Conference on Computer Vision. pp. 142–158. Springer (2020)

17. Mackiewicz, M., Berens, J., Fisher, M.: Wireless capsule endoscopy color video segmentation. IEEE Transactions on Medical Imaging **27**(12), 1769–1781 (2008)
18. Paul, S., Gundabattula, H.D., Seelamantula, C.S., Mujeeb, V., Prasad, A.: Fully-automated semantic segmentation of wireless capsule endoscopy abnormalities. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 221–224. IEEE (2020)
19. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
20. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
21. Saito, H., Aoki, T., Aoyama, K., Kato, Y., Tsuboi, A., Yamada, A., Fujishiro, M., Oka, S., Ishihara, S., Matsuda, T., et al.: Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointestinal endoscopy **92**(1), 144–151 (2020)
22. Sekuboyina, A.K., Devarakonda, S.T., Seelamantula, C.S.: A convolutional neural network approach for abnormality detection in wireless capsule endoscopy. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 1057–1060. IEEE (2017)
23. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
24. Tian, Y., Maicas, G., Pu, L.Z.C.T., Singh, R., Verjans, J.W., Carneiro, G.: Few-shot anomaly detection for polyp frames from colonoscopy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 274–284. Springer (2020)
25. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9197–9206 (2019)
26. Yang, B., Liu, C., Li, B., Jiao, J., Ye, Q.: Prototype mixture models for few-shot semantic segmentation. In: European Conference on Computer Vision. pp. 763–778. Springer (2020)
27. Yuan, Y., Wang, J., Li, B., Meng, M.Q.H.: Saliency based ulcer detection for wireless capsule endoscopy diagnosis. IEEE transactions on medical imaging **34**(10), 2046–2057 (2015)
28. Zhang, J., Zhao, C., Ni, B., Xu, M., Yang, X.: Variational few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1685–1694 (2019)
29. Zhang, R., Che, T., Ghahramani, Z., Bengio, Y., Song, Y.: Metagan: An adversarial approach to few-shot learning. NeurIPS **2**, 8 (2018)
30. Zong, B., et al.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations (2018)

5 Supplementary

Table 4. Standard performance metrics of 6-way 5-shot of the proposed technique on KID-1 dataset.

Abnormality	IOU	Dice	Sensitivity	Specificity	Accuracy
Angio	0.2804	0.3887	0.7866	0.909	0.9064
Apathe	0.1628	0.2319	0.3414	0.9415	0.892
Lymphangiectasia	0.5435	0.7025	0.9081	0.9786	0.9736
polypoid	0.1674	0.2379	0.2189	0.9637	0.8224
Bleeding	0.3289	0.4785	0.6264	0.7824	0.7476
Chylous	0.6593	0.7892	0.8125	0.9697	0.9418
Stenosis	0.3192	0.4574	0.4121	0.8797	0.7598
Ulcer	0.3550	0.4498	0.4942	0.9663	0.9451
Mean	**0.3520**	**0.4669**	**0.5750**	**0.9238**	**0.8735**

Table 5. Standard performance metrics of 6-way 1-shot of the proposed technique on KID-2 dataset.

Abnormality	IOU	Dice	Sensitivity	Specificity	Accuracy
Ampulla	0.4593	0.5917	0.8016	0.8997	0.8941
Inflammatory	0.2260	0.3222	0.4673	0.9456	0.8969
Vascular	0.1065	0.1653	0.4290	0.9317	0.9088
Mean	**0.2639**	**0.3597**	**0.5659**	**0.9256**	**0.8999**

Table 6. Standard performance metrics of 6-way 5-shot of the proposed technique on KID-2 dataset.

Abnormality	IOU	Dice	Sensitivity	Specificity	Accuracy
Ampulla	0.4646	0.5992	0.8189	0.8993	0.8952
Inflammatory	0.2275	0.324	0.4701	0.9453	0.8967
Vascular	0.1075	0.1662	0.4290	0.9138	0.9089
Mean	**0.2665**	**0.3631**	**0.5726**	**0.9194**	**0.9002**